Semantic Web Identity of Academic Organizations:

Search engine entity recognition and the sources that influence Knowledge Graph Cards in search results

Dissertation

Zur Erlangung des akademischen Grades

Doctor philosophiae (Dr. phil)

Im Fach Bibliotheks- und Informationswissenschaft

eingereicht an der

Philosophische Fakultät I

Institut für Bibliotheks und Informationswissenschaft

Humboldt Universität zu Berlin

von

Kenning Arlitsch

Die Präsidentin der Humboldt-Universität zu Berlin

Prof. Dr.-Ing. habil. Dr. Sabine Kunst

Die Dekanin der Philosophischen Fakultät I

Prof. Dr. Gabriele Metzler

Gutachter: Prof. Michael Seadle, PhD
Gutachterin: Prof. Dr. Vivien Petras

Datum der Einreichung: 2016-11-30 Datum der Promotion: 2017-01-04

Erklärung über die selbstständige Abfassung meiner Dissertation

Hiermit erkläre ich, Kenning Arlitsch, Matrikel-Nr: 563818, dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Die Dissertation wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt oder veröffentlicht.

Bozeman, Montana, USA, den 7 Januar 2017

Kenning Arlitsch

Abstract

Semantic Web Identity (SWI) characterizes an entity that has been recognized as such by search engines. The display of a Knowledge Graph Card in Google search results for an academic organization is proposed as an indicator of SWI, as it demonstrates that Google has gathered enough verifiable facts to establish the organization as an entity. This recognition may in turn improve the accuracy and relevancy of its referrals to that organization.

This dissertation presents findings from an in-depth survey of the 125 member libraries of the Association of Research Libraries (ARL). The findings show that these academic libraries are poorly represented in the structured data records that are a crucial underpinning of the Semantic Web and a significant factor in achieving SWI. Lack of SWI extends to other academic organizations, particularly those at the lower hierarchical levels of academic institutions, including colleges, departments, centers, and research institutes. A lack of SWI may affect other factors of interest to academic organizations, including ability to attract research funding, increase student enrollment, and improve institutional reputation and ranking.

This study hypothesizes that the poor state of SWI is in part the result of a failure by these organizations to populate appropriate Linked Open Data (LOD) and proprietary Semantic Web knowledge bases. The situation represents an opportunity for academic libraries to develop skills and knowledge to establish and maintain their own SWI, and to offer SWI service to other academic organizations in their institutions. The research examines the current state of SWI for ARL libraries and some other academic organizations, and describes case studies that validate the effectiveness of proposed techniques to correct the situation. It also explains new services that are being developed at the Montana State University Library to address SWI needs on its campus, which could be adapted by other academic libraries.

Zusammenfassung

Semantic Web Identity kennzeichnet den Zustand, in dem ein Unternehmen von Suchmaschinen als Solches erkannt wird. Das Abrufen einer Knowledge Graph Card in Google-Suchergebnissen für eine akademische Organisation wird als Indikator für SWI nominiert, da es zeigt, dass Google nachprüfbare Tatsachen gesammelt hat, um die Organisation als Einheit zu etablieren. Diese Anerkennung kann wiederum die Relevanz ihrer Verweisungen an diese Organisation verbessern.

Diese Dissertation stellt Ergebnisse einer Befragung der 125 Mitgliedsbibliotheken der Association of Research Libraries vor. Die Ergebnisse zeigen, dass diese Bibliotheken in den strukturierten Datensätzen, die eine wesentliche Grundlage des Semantic Web sind und Faktor bei der Erreichung der SWI sind, schlecht vertreten sind. Der Mangel an SWI erstreckt sich auf andere akademische Organisationen, insbesondere auf die unteren Hierarchieebenen von Universitäten. Ein Mangel an SWI kann andere Faktoren von Interesse für akademische Organisationen beeinflussen, einschließlich der Fähigkeit zur Gewinnung von Forschungsförderung, Immatrikulationsraten und Verbesserung des institutionellen Rankings.

Diese Studie vermutet, dass der schlechte Zustand der SWI das Ergebnis eines Versagens dieser Organisationen ist, geeignete Linked Open Data und proprietäre Semantic Web Knowledge Bases zu belegen. Die Situation stellt eine Gelegenheit für akademische Bibliotheken dar, Fähigkeiten zu entwickeln, um ihre eigene SWI zu etablieren und den anderen Organisationen in ihren Institutionen einen SWI-Service anzubieten. Die Forschung untersucht den aktuellen Stand der SWI für ARL-Bibliotheken und einige andere akademische Organisationen und beschreibt Fallstudien, die die Wirksamkeit dieser Techniken zur Verbesserung der SWI validieren. Die erklärt auch ein neues Dienstmodell der SWI-Pflege, die von anderen akademischen Bibliotheken für ihren eigenen institutionellen Kontext angepasst werden.

Dedication

For Deborah, who has inspired, loved, and supported me throughout this process, and who will be glad to have her husband back when this is over.

For my mother, Hannelore, whose love has supported me all my life.

For my librarian colleagues: there is great work ahead of us.

Acknowledgements

I have written in Chapter 1 that the origins of this research can be traced to 2010, but the groundwork was laid much earlier, perhaps even dating to an unsatisfying undergraduate experience that lacked the passion that arises when a problem must be solved. Robert Pirsig described this lack in an educational system that was designed as a passport to a career rather than a passage to enlightenment, in a book I read in 1987 and whose lessons remain with me still (Pirsig 1974). I considered a doctorate after I earned my master degree in 1993, but life took its twists and turns and it wasn't until 2010 that I began to find the research interests and the personal and professional influences I needed to begin this course of study.

Top among those influencers is my wife, Dr. Deborah Keil. Without the love and stability she brought to my life, none of this would be possible. She is my Dr. Awesome.

My advisor, Dr. Michael Seadle willingly took me on as his student when I first approached him in 2013 and has been generous in his advice. I am grateful to him for this opportunity to continue my education. Dr. Vivien Petras served as my second reader, and her close reading of my drafts and many comments pushed me to improve when I thought I was nearly finished.

My research partner since 2010, Patrick OBrien, has brought a wealth of ideas and expertise since we began working together that has stimulated my research, and I look forward to further collaboration.

Andrew Richardson, who has been my friend since junior high school, fostered a technical path for me by building my first computer just as I entered library school in 1992. His timing couldn't have been better, as the Web was about to explode onto the scene.

Dr. Gregory Zick has long been my mentor and friend, and he was a catalyst for my early digital projects that eventually led to this line of research.

I am grateful to Dale Askey, my colleague and friend, for suggesting that it might be possible for me to earn my doctorate at Humboldt Universität zu Berlin, and for volunteering the *McMaster University Library* as one of the test cases described in this dissertation. I look forward to cheering him on as he begins work on his own dissertation.

Dr. Stephanie Krueger, made it to the Ph.D. finish line before I did and then kindly turned around to help me navigate the process with valuable advice. Dr. Clifford Lynch, Executive Director of the *Coalition for Networked Information* was generous in allowing me to use CNI as a test case, and Diane Goldenberg-Hart and Maurice Angelo Cruz were extremely helpful as we struggled to right the SWI ship for CNI. The editing prowess of Dr. Jennifer Askey was instrumental to the final stages of writing, and she pushed me to further develop and clarify some nascent ideas.

In 2012, Deborah and I moved to Bozeman, Montana, so that I could begin my first position as a library dean at Montana State University and she could take her place among the faculty of the Department of Microbiology and Immunology. Our lives and careers have blossomed at MSU, and we are both so appreciative of our community of colleagues and friends. My faculty and staff at the library are among the brightest and most collegial that I have seen anywhere, and my administrative team has been patient with me as I struggled to complete this work in addition to my full-time job. I also could not have hoped for a more inspiring and supportive group of colleagues in my fellow deans and the vice presidents. The former provost at MSU, Dr. Martha Potvin, hired me and then supported without hesitation my aspiration to earn my doctorate, and Dr. Robert Mokwa, interim provost at Montana State University, has continued Dr. Potvin's support. President Waded Cruzado has set a tone and expectations at MSU that have propelled excellence and growth, and I am proud and thankful to be a part of the leadership team.

Dr. Lillian Lin leads the *Statistical Consulting and Research Services* group at Montana State University; without her and her students I could not have conducted the statistical analysis that is evident in this dissertation. One of those students, the (now) Dr. Katherine Banner, was a great teacher, and I received additional advice from Claire Rasmussen and Jordan Schupbach.

Justin Shanks, Montana State University's Semantic Web Identity Researcher since August 2015, quickly grasped the concept of SWI and has developed the service at MSU Library beyond what I had envisioned. Though busy with his own dissertation he kindly offered to read my first draft and gave me valuable comments.

Finally, thank you to Ted Stazeski, for friendship, encouragement, and some of my best days of desert wanderings. May you continue to thrive on the bubble of the cusp of the vanguard of the bleeding edge. Way out there.

Table of Contents

List of Tables

List of Figures

List of Equations

Glossary of Abbreviations and Terms

ARL	Association of Research Libraries
CMS	Content Management System
DAM	Digital Asset Management
DBpedia	A knowledge base that extracts structured content from Wikipedia articles and makes it freely available on the Web as linked data.
Google+	Google Plus
GMB	Google My Business
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
JSON	JavaScript Object Notation
KC	Knowledge Graph Card
LIS	Library and Information Science
LOD	Linked Open Data
MSU	Montana State University
PDF	Portable Document Format
PNG	Portable Network Graphics
R	R statistical software package
RStudio	Development environment for R
RDF	Resource Description Framework
Schema.org	A shared vocabulary for structuring data on the Internet. It is supported by the major search engines (Google, Bing, Yahoo!, Yandex).
SEO	Search Engine Optimization
SEM	Search Engine Marketing
SERP	Search Engine Results Pages
SWI	Semantic Web Identity
SWIR	Semantic Web Identity Researcher

SWO	Semantic Web Optimization
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
Wikidata	Project of the Wikimedia Foundation that allows community creation of structured data records for Wikipedia articles.
Wikipedia	Online community-developed encyclopedia, created and managed by the Wikimedia Foundation
YAGO	Yet Another Great Ontology

Chapter 1 Introduction

Section 1.1 Problem Statement

The World-Wide-Web that was launched in the early 1990s revolutionized the Internet. Built on the HyperText Transfer Protocol (HTTP) and graphical browsers, the "Web" quickly became a fixture of popular culture and business processes, ushering in an era of tremendous technological growth (Bryant 2011) and forever "changing the way we work, socialize, create and share information" (Manyika and Roxburgh 2011). The Web disrupted many industries, rewarding businesses that saw its potential and adjusted appropriately, and leaving behind those that did not (Christensen 1997). It catalyzed technological and cultural change in libraries by flipping the existing paradigm of information scarcity to one of information abundance, (Dempsey 2014), a change that was perhaps felt most acutely in the domain of reference services (Kennedy 2011) as a vast array of information sources became available to users without the need for librarian mediation. Few industries, for profit or not, were untouched. The travel industry was also transformed, leaving the number of travel service agencies greatly diminished as customers were able to satisfy their booking needs independently. Although the Web disrupted many industries, it also spawned massive new economic and technological development, giving rise to the "dot com" boom and tens of thousands of new businesses. The U.S. employment sector that supplies, stores, and provides access to information, alone, increased from 60,000 jobs in 1990 to over 260,000 jobs in 2016 (Bureau of Labor Statistics 2016). A 2011 report estimated that the "Internet accounted for 21 percent of GDP growth over the last five years among the developed countries" (du Rausas et al. 2011). The impact of the Web on the growth of the overall Internet has been undeniable.

The innovative feature of the Web in its first iteration (Web 1.0) was its ability to link documents to each other using Uniform Resource Locators (URLs) inserted into the text of documents written in the HyperText Markup Language (HTML). Explosive growth of documents posted on thousands, then millions of websites quickly led to a disorganized state in which it was difficult to find anything. The initial solutions to the increasing information organization problem were akin to those employed by librarians for generations, i.e. manually-created subject directories, such as the *Yahoo Directory; DMOZ;*

Scout Report; and the Librarians' Index to the Internet. Nearly all of these directories, including the *Yahoo Directory*, which was once "the most common way people found websites" (Sullivan 2014) are now defunct because manual organization by humans simply isn't feasible in the dynamic environment of the Web. The deluge of information requires the processing power of machines.

In the wake of failed subject directories, search engines found greater success with algorithms that weigh numerous "signals" to return accurate results to search queries. The search engine landscape was crowded in the 1990s and early 2000s ("Timeline of Web Search Engines" 2016), but the "PageRank" ranking algorithm developed by Sergey Brin and Larry Page (Brin and Page 1998) that launched the Google search engine proved superior to other models. Google has since has become the dominant public search engine on the Internet, consistently controlling nearly two-thirds of the United States search engine market. Microsoft's Bing search engine, which also provides organic search results for Yahoo! is the only other significant contender in the United States (comScore, Inc. 2016). Google's market dominance in the countries of the European Union is said to exceed 90% (Meyer 2015).

Long before the size of the "indexed" World-Wide-Web rose to nearly 5 billion pages (de Kunder 2016), search engine and Web developers began to realize that the "strings-based" search environment of the Web, where algorithms matched strings of text to search queries, was a limited solution. The Web had to evolve into a new environment where the subtleties of context and meaning could be gleaned by machines, and where search engines were more likely to provide answers to queries rather than simple referrals to documents where the answers might reside. Thus, the idea for the Semantic Web was born (Berners-Lee, Hendler, and Lassila 2001). Further explanation of the evolution and structure of the Semantic Web will be addressed in Chapter 2.

In the still-developing environment of the Semantic Web, search engines seek to establish facts about "entities," which are defined as people, places, organizations, landmarks, etc. For the purposes of this dissertation entities will be defined as academic organizations, which in some academic environments may be known as "units" in the organizational hierarchy. Search engines may establish facts organically as they find mention of entities while indexing documents at websites. But facts may also be established proactively by the entities themselves, in knowledge bases established for this purpose. In

this dissertation, the author explores and tests the proactive approach, which appears to be lacking in most academic organizations.

The facts gathered by search engines are largely invisible to the public, as they reside behind the scenes in graph databases. But one visible manifestation of these facts is the Knowledge Graph Card (KC) that Google began to display in 2012 (Singhal 2012) in its search results, and Bing followed a short time later with its own version of the KC. Aside from providing quick and easy "answers" about an organization to users, the author posits that the KC may be viewed as an indicator that the search engine has discovered sufficient verified facts about the organization to establish it as an entity in its knowledge graph. The author characterizes this condition as Semantic Web Identity (SWI). Conversely, when a KC fails to appear for an organization or it displays few or inaccurate facts, the condition is characterized by this author as lacking or poor SWI.

The study conducted in this dissertation measures the existence of KC for the 125 member libraries of the Association of Research Libraries (ARL). It also measures the number of facts (information elements) displayed on the KC as an indication of KC "robustness," i.e. the more facts that are displayed, the more robust is the information the search engine has about the organization. Finally, the research measures the presence of structured data records for the organizations in certain Semantic Web knowledge bases to determine whether the lack of presence in those knowledge bases correlates with a lack of KC.

Findings of the research demonstrate that many ARL member libraries suffer from poor SWI, as measured by the presence of accurate and robust KC. Most of the libraries that have poor SWI usually also lack structured data records in the surveyed knowledge bases, which the author hypothesizes are sources of verified data about entities that Google uses to populate its Knowledge Graph, and from which it then generates a KC.

While the survey of ARL libraries constitutes the data set for this dissertation, it is clear from more informal surveys conducted by the author that lack of SWI extends to other organizations in academic institutions. Search engines have difficulty realizing and verifying the existence and the nature of organizations across the academic hierarchy: colleges, departments, centers and institutes.

In the world of academia, SWI is generally most robust at the top levels of the institution. Searching for a university will usually yield search engine results pages (SERP)

that imply machine comprehension of the institution, as indicated by the presence of a robust KC. That comprehension diminishes as the search moves deeper into the organization to the level of college, department, institute and center. As will be demonstrated in this study, those organizations often lack KC, display KC with a paucity of information, or even display KC that show an organization that is different from the one searched (see Figure 1). Yet it is on the websites of these organizations where activities and products of interest to students, researchers, faculty colleagues, and funders of the institution are most likely to be expressed. A lack of comprehension of these organizations can logically be assumed to result in fewer referrals; if the search engine doesn't understand the organization or trusts that it will provide a good experience, then it is less likely to send its users there.

Information-seeking behavior in this age is almost always centered on the use of Internet search engines (Connaway and Dickey 2010), but the effect of SWI in the discovery process is only gradually becoming understood. Lack of SWI, or incomplete or inaccurate SWI for organizations may result in search engines failing to refer users whose interests match those organizations. That, in turn, may affect the parent academic institutions' abilities to attract research funding, faculty talent, and students. For instance, funding agencies that seek evidence that a university is engaged in specific research may fail to find a credible connection if the relevant research center hasn't addressed its SWI. Students seeking a match for their study interests may likewise not be referred if a particular university department hasn't addressed its SWI, leaving search engines to rely on potentially inaccurate information that other sources may supply (DePianto 2016). These interrelated factors could negatively affect University rankings and reputation if research funding doesn't find its way to the institution, if student enrollment declines, or if faculty can't be recruited because they are not attracted by the university's reputation in their research interests.

The research described in this dissertation is supplemented by the description of a process developed at Montana State University that academic libraries could use to improve SWI for themselves as well as for other organizations at their institutions. This process involves populating appropriate Linked Open Data (LOD) and proprietary knowledge bases with accurate and verifiable information. A positive effect on the SWI of academic organizations can be expected from these actions, but determining which knowledge bases

are most effective at populating search engine knowledge graphs is difficult because commercial search engines like Google and Bing are notoriously secretive about their methods, often revealing only fragments of information that may be used for guidance. The secretive nature of these search engine companies stems from the competitiveness of the commercial world. Revealing too much about proprietary systems can lead competitors to copy and improve methods, and it can also lead to the application of "black hat" SEO practices to unfairly advance products in SERP. The competition among search engine companies may also lead them to frequently change or improve their methods, rendering specific pathways obsolete.

With these limitations in mind, the processes described in this research are aimed at being indicative rather than prescriptive. Which knowledge bases should be populated is not as important as an overall awareness by librarians of the growing importance of LOD and other sources from which search engines may draw to build their knowledge graphs. While the research in this dissertation focuses on Google and its related products, the concepts should be adaptable as other semantic search engines develop their use of knowledge graphs.

There is no evidence in the library and information science (LIS) literature to suggest an awareness of SWI. Moreover, LIS programs are not currently teaching the steps required to correct the condition, nor are academic institutions implementing those steps in any systematic manner. The problem of SWI overlaps the fields of LIS and marketing, revealing the awkwardness of traditional marketing methods in a machine-based environment, where inconsistent branding and terminology, as well as failure to engage with appropriate data sources, can have widespread consequences.

The need to establish SWI for academic organizations represents an opportunity for the library profession, whose mission has always included the creation and maintenance of structured data records. Academic libraries could develop formal services to create and maintain these records for other organizations on their campuses, and these services could be viewed as core to the teaching, research and outreach missions central to so many universities. Most libraries will be challenged to develop these services because it forces them to move outside their normal sphere of operations and into the Semantic Web, where few librarians have expertise. A commitment to learning new skills will be required to accompany an overall strategy.

Section 1.2 Research Background

The field of search engine optimization (SEO) became significant for the author in 2009, when, after a decade leading the development of the digital library at the University of Utah, he realized that most of the objects he and his team had loaded into digital repositories were not visible through popular Internet search engines. Further investigation revealed that the problem was widespread, and that few libraries were successful in getting their digital objects to appear in SERP. Several years of funded research followed, during which the author and his research partner presented and published widely on the SEO problems of digital collections, the related problem of SEO for institutional repositories, and solutions, which were not always technical in nature (Arlitsch, OBrien, and Rossmann 2013; Arlitsch and OBrien 2013; Arlitsch and O'Brien 2012).

The related phenomenon of SWI was first realized by the author in the fall of 2012, when he began work as dean of the library at Montana State University. While conducting a Google search for "Montana State University Library," he observed a KC in the search results that displayed information for a branch campus library in Billings, MT, rather than the main campus library in Bozeman (Figure 1).

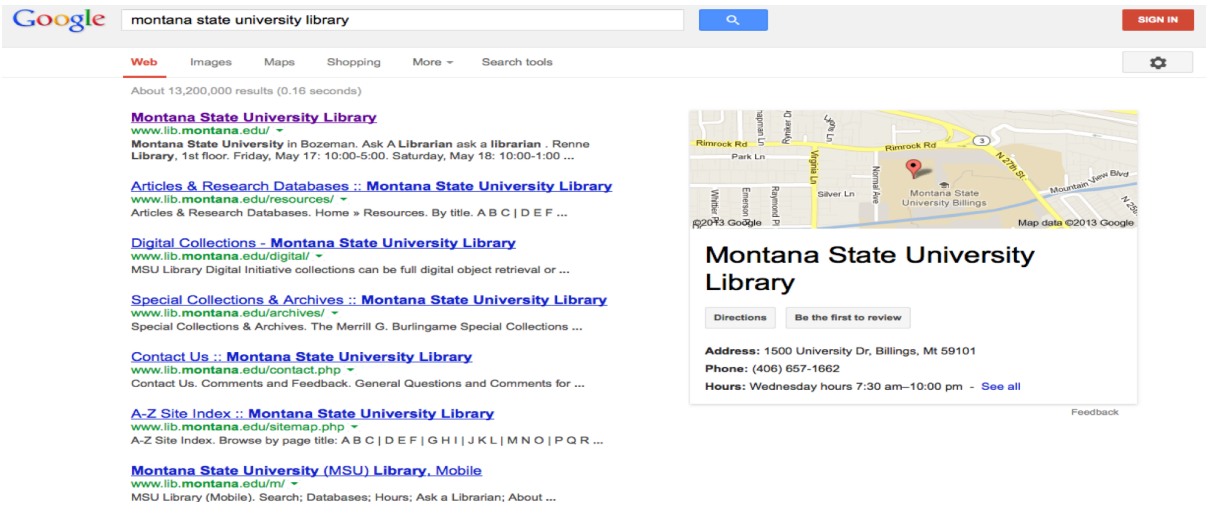

Figure 1: Montana State University Library KC in 2012

The author and his research team began to investigate the reasons for the inaccurate KC and gradually learned to correct it so that it displayed accurately for the Montana State University Library and displayed more verified facts about the organization

(Figure 2). A broader investigation led to the realization that inaccurate or incomplete KC was a widespread problem, not only among libraries, but also across other academic organizations. The author began gathering evidence of this problem and launched a formal course of study with the *Institute für Bibliotheks- und Informationswissenschaft (IBI)* at *Humboldt Universität zu Berlin* in 2014, resulting in this dissertation.

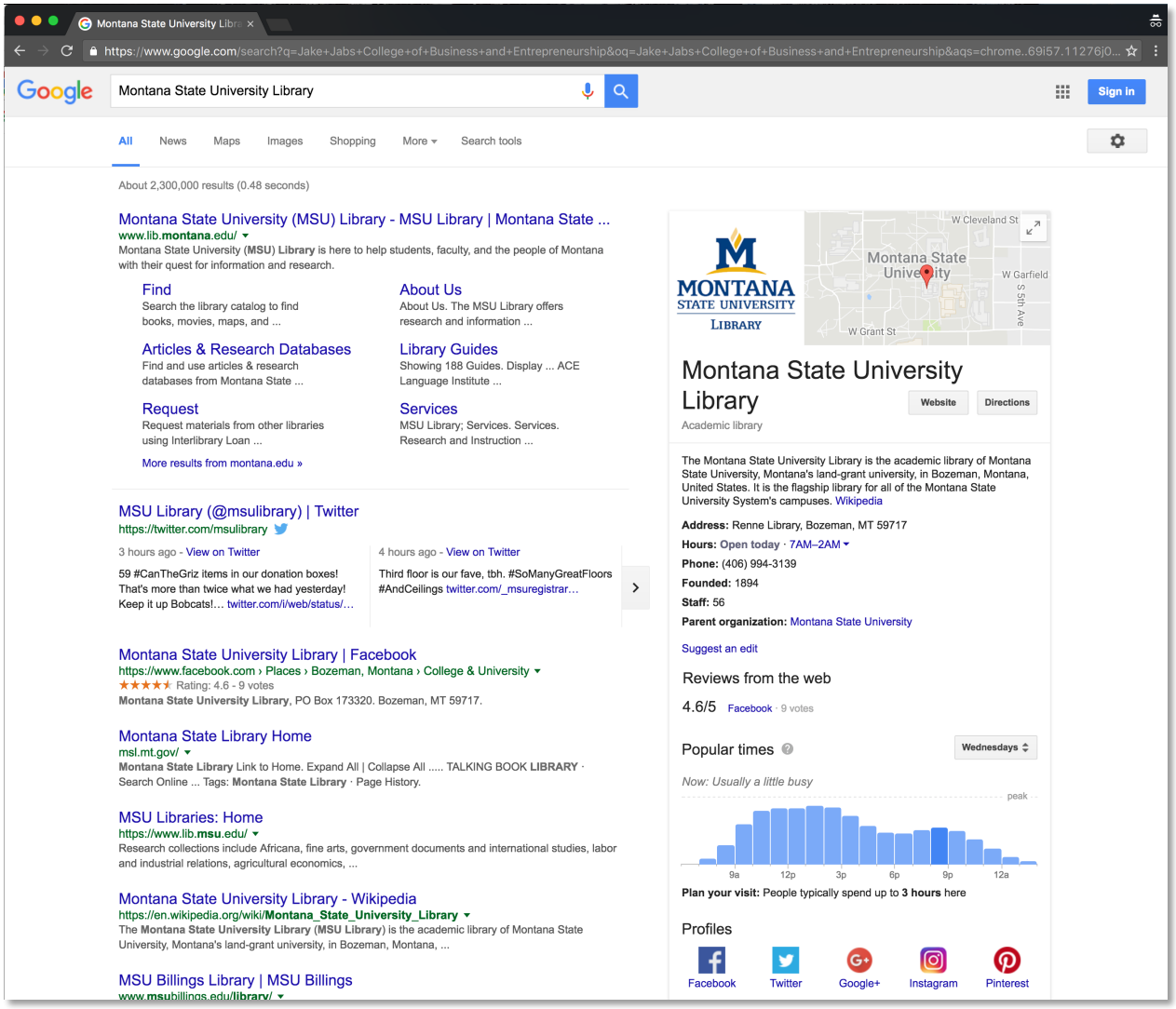

Figure 2: Montana State University Library KC in 2016

The promise of search engines that are fully engaged with the Semantic Web is that they will deliver more accurate and relevant results, including enhancements such as the KC. Search engine developers anticipate that users who enter a query for an organization want to know about that organization, but may also want to locate or contact the organization; hence the possible "information elements" that may be displayed on the KC. These

elements may include, but are not limited to: address; telephone number; hours of operation; organization type; description; a link to the website; and a link that can direct the user from his/her current location via a mapping application. In short, the structured and contextual data of the Semantic Web creates the potential for a better user experience by anticipating what searchers seek, and by connecting them more effectively with the object of that search. Search engines that are engaged with Semantic Web data sources are more capable of delivering answers, rather than the traditional list of links that are expected to be relevant to the search. Beyond search engines, Semantic Web data records are also utilized by a variety of semantic technologies in mobile and other devices, enabling a host of location-aware social, educational and industrial applications (Bizer et al. 2009). Mobile and desktop applications such as Google Maps and voice-activated answer applications such as Apple's Siri are two examples of semantic technologies, and there is potential to develop a spectrum of applications that use data from sensors. Linked data coupled with sensors placed in areas around a city, for example, creates the possibility of applications asking for "parking spots in Berlin to calculate the city's availability of car parking places" (Pfisterer et al. 2011).

Google and Microsoft have been developing knowledge graphs (generically known as graph databases) for some time, and those knowledge graphs now significantly inform Google and Bing search engines, helping them to "create interconnected search results that are more accurate and relevant" (Patel 2015). Google's Knowledge Graph and the more recent Bing Satori may be populated by facts that the search engine's crawlers can gather organically, simply by combing through websites and applying algorithms to establish relationships. However, more verifiable facts are harvested from data sources in the Linked Open Data (LOD) cloud as well as proprietary sources (Dame 2015). The LOD cloud, which began with 12 data sets in 2007 (Lalithsena et al. 2013) and has grown to over 1,000 in 2014 (Schmachtenberg, Bizer, and Paulheim 2014), "reveals the interests of data publishers to publish their data as structured data" (Lalithsena et al. 2013). Data sets published in the LOD are freely available to anyone.

Academic library organizations suffer from poor SWI at least as much as their sister organizations in academic institutions. During initial exploration of this topic in December 2014, the author conducted an informal precursor survey of the 125 members of the Association of Research Libraries; it revealed that many of the libraries displayed no KC in

search results at all, while many of the remainder showed very poor KC. The historic reluctance of many librarians to engage in new Internet data tools and sources such as Google and Wikipedia have not served the profession well; librarians' slow pace of adoption and skill development has almost certainly contributed to the current state of SWI for academic libraries. However, the situation also presents new opportunities for libraries to develop skills and services to help themselves and their institutions in very measurable ways.

SEO has long been an accepted practice to gain visibility and rankings in SERP. Semantic Web Optimization (SWO) may be considered an extension of SEO as it continues the effort to improve relationships with search engines in the Semantic Web environment. In the context of this dissertation SWI is a goal of SWO. The tactics of SWO differ from SEO in that they rely more on engagement with LOD and proprietary data sources that search engines trust and understand, as well as creating structured data markup for websites themselves. Although on-page Schema.org markup in websites may also be useful in establishing SWI, Schema.org is not a focus of this dissertation.

Section 1.3 Research Hypothesis

This research tests the hypothesis that certain proprietary and open Semantic Web knowledge bases must be actively engaged so that an organization can be optimally recognized and understood as an entity by Google's main search engine. A review of the literature (detailed in the next chapter) as well as the author's own testing has led the research described in this dissertation to focus on five knowledge bases as potential sources of information that help populate Google's Knowledge Graph, which in turn generates a KC in Google SERP. These five knowledge bases are: Google My Business, Google+, Wikipedia, DBpedia, and Wikidata.

Active engagement with these five knowledge bases is defined as someone having created or improved records that represent the organizations in question. The presence of records in the knowledge bases (or lack thereof) will be observed for the ARL member libraries and compared to the observation of accurate KC for the same member libraries in Google SERP. It is anticipated that organizations with records in these knowledge bases are

more likely to display accurate and robust KC. The organization can be said to have achieved SWI if a KC appears that includes accurate facts about the organization.

The findings will show that academic libraries represent themselves inconsistently on the Semantic Web, a problem that looms large in the research. This phenomenon will be explained in greater detail in chapters 3, 4, and 5, but for now, it is sufficient to note that most ARL libraries are known by more than one name, and the author refers to these as primary and alternate names.

Section 1.3.1 Research Goals

The topic of this dissertation includes elements of ethnographic and computer science build-and-test approaches, but it does not fall exclusively into either realm. Rather, the dissertation uses a hybrid research methodology known as "action research," which was developed in the social sciences, but has also been widely used in the fields of business and information systems, as well as library and information science. Action research involves a cyclical process of gathering and analyzing data, reporting results, and then designing an action plan and evaluating the results of that action. Data that the author gathered for this research include screen captures from SERP and data records from the five LOD and proprietary knowledge bases named earlier. The Action Research Methodology section in Chapter 2 will provide more background and detail on the action research process, and Chapter 3 will show how the research methods follow the methodology.

An initial goal of the research is to demonstrate the current condition of SWI among the 125 member libraries of the Association of Research Libraries (ARL). Since 94 of the libraries also represent themselves with a second (alternate) name, a total of 219 names will be searched in Google and results of the SERP will be recorded for each to show whether a KC appears and whether it is accurate for the organization being searched. Robustness will be scored by the presence of certain information elements shown on the KC. The five knowledge bases listed in Section 1.3 will be searched to determine whether records exist for both names of the libraries. An effort will be made, using logistic regression, to predict which knowledge bases are most likely to influence the appearance and robustness of the KC.

Beyond showing the results of the ARL dataset, the research will also provide evidence that SWI concerns extend to other academic organizations by showing examples from Montana State University. Additionally, three case studies will demonstrate an evolving process that has proven successful at establishing SWI. Finally, a new service at Montana State University will be described, whose intention is to help academic organizations achieve SWI.

Section 1.3.2 Research Questions

Research Question 1: What is the current state of Semantic Web Identity of ARL libraries, as indicated by the presence of accurate Knowledge Graph Cards in Google search results when the primary and alternate names of those libraries are searched?

Research Question 2: Are records or profiles present for ARL primary and alternate library names in the following knowledge bases: Google My Business, Google+, Wikipedia, DBpedia and Wikidata?

Sub-question 1: Is an accurate KC likely to display in search results if the library organization has not been claimed and verified in Google My Business?

Sub-question 2: Is a KC likely to display a description field (one possible information element) if no Wikipedia article exists for the primary or alternate name of the library?

Research Question 3: Does the presence of a given knowledge base record predict the odds of other facts (information elements) on the KC being populated?

Section 1.4 Structure of the Dissertation

Chapter 2 examines the scholarly context for the research by explaining the evolution of the topic from its SEO roots and its relationship to the more recent practice of Semantic Web Optimization (SWO). This chapter includes a review of the literature for SEO, Search Engine Marketing (SEM), SWO, knowledge graphs in search engines, and SWI, as well as tracing the development and applicability of the action research methodology.

Chapter 3 details the specific research methods based on formal steps in the action research methodology. First, there is an outline of steps taken to gather and analyze data from a national sample of 125 ARL libraries, including searches of their primary and

alternate names, plus a smaller sample of more general academic organizations at Montana State University. Second, specific actions are described that were taken to establish or improve SWI in case studies involving two academic libraries and one professional organization. Finally, the chapter discusses the limitations of the research methods.

Chapter 4 describes the findings of the data gathered from the survey of ARL libraries, including descriptions and examples of the statistical analysis equations. It also details the changes that were made to effect SWI for the three organizations that served as case studies, and briefly describes the results. The chapter concludes with a brief description of findings for eleven MSU colleges.

Chapter 5 discusses the findings in greater detail, including examples of specific findings from certain ARL libraries. The author also elaborates on the five knowledge bases and speculates on the impact of the primary and alternate names.

Chapter 6 explains the broader implication of the research, including the SWI services that MSU Library is developing, which other academic libraries could adapt for their constituents.

Chapter 7 concludes the dissertation with a final discussion of the usefulness of this study, and the future steps might be taken with the data set.

Chapter 2 Scholarly Context

Section 2.1 Introduction

This research shows that academic organizations are poorly represented on the Semantic Web because search engines often do not recognize that those organizations exist, where they are located, or the nature of their businesses. Semantic Web Identity (SWI) for those organizations cannot be established because the knowledge graphs that increasingly inform search engines seem to lack enough verifiable facts to establish the organizations as entities with relationships to other entities. Academic organizations have contributed to their poor SWI because they have not proactively engaged in the sources that influence search engine knowledge graphs, and because they represent themselves inconsistently, causing confusion in the machine-based environment of the Semantic Web.

Semantic Web Identity (SWI) may be considered a form of marketing because it is so intertwined with the way organizations represent and promote themselves, and while traditional marketing goals are still valid, the environment of the Semantic Web substantially alters the application of marketing practices to achieve those goals. This chapter will begin by reviewing the literature for traditional marketing practices in academic libraries before moving into a review of how libraries approach marketing on the Semantic Web. The steps to attain SWI are related to the long-standing practices of Search Engine Optimization (SEO), and therefore a review of SEO is appropriate. This will be followed by the description of a newer area of SEO called Semantic Web Optimization (SWO). The literature review will continue with an explanation of Semantic Web Identity (SWI), the term coined by the author to describe the focus of this dissertation, and will review the development of Knowledge Graph Cards (KC) by search engines companies. The chapter will conclude with a review of the action research methodology, describing its origins, the disciplines that have found it useful, and its value for the research in this dissertation.

Section 2.2 Marketing in Academic Libraries

The names by which academic libraries choose to represent themselves are crucial to the creation and maintenance of SWI. The consistent application of a name or brand is a fundamental marketing strategy, but it is one with which academic libraries seem to

struggle. Most academic libraries have what can be considered primary (official) names that are based on the institution to which they belong, e.g. “University of Utah Library.” However, most member libraries of the Association for Research Libraries (ARL) also have alternate (unofficial) names for their libraries, which were commonly designated after substantial donations to the library. Following the example above, the alternate name of the “University of Utah Library” is the “J. Willard Marriott Library,” after the founder of the Marriott hotel business, who donated a sum of money when the new library building was completed in the mid 1960s. For decades, academic libraries have used the primary name of their organization in some venues and the alternate name in others. In an analog environment, this inconsistent use of names, while a poor marketing practice, is less problematic than in a digital environment because humans are very good at making mental connections to understand variant terms for entities. Machines, on the other hand, do not have this capacity unless they have access to data records that explicitly establish the relationships of the names for them. The growing field of artificial intelligence notwithstanding, for the foreseeable future the machines that support search engines must be explicitly informed of the relationships of name variations.

Definitions of marketing have changed over time, but in recent years the scholarly community has come to general agreement that marketing is “the strategic business function that creates value by stimulating, facilitating and fulfilling customer demand” (Palmer 2009). The terms “marketing” and “branding” are used interchangeably in some of the literature, but there is a distinction. Some describe a brand as “a cluster of functional and emotional values, which promise a particular experience” (de Chernatony 2002), while others espouse a more philosophical view: “Branding is the expression of the essential truth or value of an organization, product, or service” (Heaton 2011). While marketing is considered a “push” activity that includes a variety of strategies and tactics, branding is what differentiates products, organizations and services; it is considered a “pull” operation that develops loyal customers. The establishment of product and organizational names as part of the branding process has been studied widely in the business literature (Kohli and LaBahn 1995; Rooney 1995), which emphasizes consistent use once the brand is established (*The Economist* 1988).

The LIS literature that covers marketing rarely addresses the problem of inconsistent use of names or brands on the Semantic Web. Most of the marketing literature in LIS is

framed as "outreach," and is limited to discussions about traditional services and print-based outlets (Dennis 2012; Fabian et al. 2003; Carter and Seaman 2011), although more recent publications have emphasized and evaluated the use of digital social media networks as mechanisms for library outreach (Young and Rossmann 2015; Vucovich et al. 2013; Alkindi and Al-Suqri 2013). Singh noted a lack of marketing culture in libraries, stating that "branding has yet to receive its due consideration in library and information services" (Singh 2004). When brand names are discussed it is usually as a promotional strategy for the website and the library databases listed there (Hepburn and Lewis 2008); while this particular article mentions search engines in the introduction there is no further discussion of how brand names might interact with them. Rowley noted long ago that "library Web sites have been preoccupied with Web site functionality, and have not lingered long on the question of brand or corporate identity" (Rowley 2004). More recently, she and a co-author stated that "Brand consistency is central to ensuring brand impact and the building of brand equity" (Rowley and Edmundson-Bird 2013). Despite their article's focus on branding in digital spaces, it does not delve into any significant discussion of the importance of consistency of library names.

One specific mention of inconsistent representation of library names appears in an article about libraries' engagement with Pinterest, where the authors found that 39 percent of libraries surveyed "failed to display the complete library name and institution (college or university) name on the profile title or profile picture…In most cases, institutional identities could be determined only by following the link provided to the institution's library Web site" (Thornton 2012). Another article describes a survey of over 700 libraries that explores the history of naming academic library organizations, buildings, or collections, concluding that "there seems to be no consistent naming practice" and even that the "concept of libraries as opposed to collections [that may have been named after individuals] is not consistent" (Crosetto and Atwood 2012). These two articles were the only ones found in this literature review that specifically mentioned the problems of inconsistent use of library names, and even they did not delve very deeply. More recent publications address the use of websites as marketing outlets, but they make only passing reference to the importance of naming (Higginbottom and Gordon 2016), and none to marketing strategies specifically for the environment of the Semantic Web.

A recent survey compares strategic plans in academic libraries to the top LIS trends defined by *Association of College & Research Libraries (ACRL)*, *The Horizon Report*, and *Ithaka S+R*. Marketing and outreach ranked high on the list of top trends, and the author reports that "71.4%" [of] plans included marketing, public relations, or a similar form of outreach as one of their goals" (Saunders 2015). But although 22% "specifically mentioned use of social media and networking tools such as Facebook, Twitter, and blogs as marketing or communication venues," any mentions of brand consistency or feeding structured knowledge bases that might help populate search engine knowledge graphs are conspicuously absent.

In summary, while marketing literature in the business world discusses the value of name brands and their consistent use, the LIS literature is almost devoid of any similar discussion.

Section 2.2.1 Search Engine Marketing (SEM)

Search Engine Marketing (SEM) is related to both marketing and SEO but is typically defined as increasing website visibility in SERP through paid advertising ("Search Engine Marketing" 2016). Ninety percent of Google's $75 billion revenue in 2015 was earned through its "Pay Per Click" advertising model (Alphabet, Inc. 2015). The business management literature is filled with articles about SEM strategy (Sen 2005; Dou et al. 2010; Skiera, Eckert, and Hinz 2010; Panda 2013), but here again, references to consistent naming practices are difficult to find. It appears the business world may also be slow in recognizing the need to transition analog marketing practices to the Semantic Web.

Libraries typically do not participate in paid advertising on the Web and therefore SEM is not a directly viable channel. However, libraries can take advantage of some of the powerful tools that Google has developed for advertisers and use them to develop marketing tactics. For instance, the Keyword Planner in Google AdWords (Google, Inc. 2016b) can help a library determine which words or phrases are searched more frequently by users and are therefore more likely to drive traffic. In the Keyword Planner, one can find that the phrase "university library" is searched more frequently by users than "academic library" and the phrase "academic papers" is searched far more often than "institutional repository." This information can be useful to libraries as they develop their websites and

plan their SWI, as the terms can be utilized in records or articles in numerous knowledge bases.

Section 2.3 Search Engine Optimization

The search engine business market is dynamic and competitive. For as long as search engines have been a force on the Internet, organizations of every kind have practiced techniques that would help them and their products become visible and highly ranked in SERP. “Whether a company has products to sell or is simply trying to achieve a wider audience or more page hits, the obvious way to achieve its goal is to appear in the place where most of the searcher’s attention is focused” (Cahill and Chalut 2009). SEO is the practice of various techniques that help achieve this visibility, and priorities focus on three major steps. The first step assures that digital objects will appear in a search engine’s index, the result of which is sometimes referred to as “indexing ratio;” the second step aims to achieve a high ranking in SERP; and the third improves relevance of search results to users by providing descriptions (i.e, “rich snippets” that appear under the links) that will increase click-through rates (Arlitsch 2015).

In industry, the difference between products appearing on the first page of SERP or several pages deep can mean success or failure of the business. In the traditionally subsidized world of academia the fallout from poor performance in search engines is less dire, but even in this environment the pressure to show return on investment (ROI) has been mounting (Johnson et al. 2014; Nickolai, Hoffman, and Trautner 2012). A successful SEO program can connect a university’s Web properties to its intellectual output, helping demonstrate the impact an academic institution has on its students, the research world, and its community. The commitment to SEO success in both industry and academia has been mixed, depending on available expertise and resources. Large business organizations are most likely to be able to hire experts from the plethora of SEO consulting firms, but money doesn’t necessarily promise success. The pressure to deliver results in the business world has sometimes led to “black hat” SEO practices that have resulted in search engines banning sites from their indexes (Segal 2011).

Search engine indexes are populated by software applications known as “crawlers, spiders, or bots” that navigate through websites by following links, and then harvesting the

text that is displayed on web pages. Static HTML pages displaying digital objects were common in the early days of the Web and seemed easiest for crawlers to harvest, but generating and managing static pages was not a scalable practice. Thus, database-generated websites became common in the early 2000s, using platforms commonly known as Content Management Systems (CMS). Digital Asset Management (DAM) systems are a subset of CMS and are used to manage large numbers of digitized objects and their metadata. Libraries began to use DAM-driven repositories extensively in the early 2000s, but those repositories have sometimes struggled for inclusion in Internet search engine indexes. Many repositories continue to suffer from low indexing ratios, meaning that only a small percentage of their digital objects' metadata have been harvested by crawlers and added to the search engine's index (Arlitsch and OBrien 2013). As a result, relatively little traffic is directed to library sites from the billions of queries submitted to search engines each month (comScore, Inc. 2016). The reasons for this can be divided into three broad categories:

1. Technical – Common technical barriers have included issues around hardware and software, website design, and metadata. Data interchange standards developed by libraries over decades (MARC, EAD, TEI, Dublin Core, OAI-PMH) were developed with little consideration for integration with Internet search engines (Arlitsch 2014a).
2. Organizational - Few libraries have implemented holistic and strategically-oriented search engine optimization programs (Arlitsch, OBrien, and Rossmann 2013). Many leave SEO to their IT departments because SEO is viewed as a purely technical issue and because few library administrators understand it. This is a strategic error: "an IT department should not be left to make … the choices that determine the impact of IT on a company's business strategy" (Ross and Weill 2002).
3. Cultural – Many librarians disparaged Internet search engines when they were first developed, and avoided using or teaching them (MacColl 2006). Some practiced "public derogation coupled with private adoption" (Anderson 2005) wherein librarians advised their students and other users to beware of the Google search engine while often reaching for it as a first resource, themselves. Similarly, librarians often disparaged Wikipedia (Luyt et al. 2010) in its initial years and were slow to engage, and while acceptance has improved, widespread understanding of its role as a data source for the Semantic Web is still lacking.

The complexity of DAM systems and a lack of awareness about SEO from DAM developers often created barriers to search engine crawlers, including multiple (non-canonical) links and labyrinthine paths to objects. Contrary to popular belief, search engine crawlers do not "crawl" the contents of a database. Instead, they trigger "clicks" of hyperlinks on websites to generate the HTML pages that are compiled from various database elements, and then they harvest the indexable text that is displayed (Arlitsch and OBrien 2013). Some authors have suggested replicating static pages outside the DAM as a way to respond to these problems: "Unless links are located on a static Web page, crawlers won't find them" (DeRidder 2008). Few modern websites, though, have returned to static pages because that method is not scalable or manageable.

Other specific technical barriers to SEO can include problems with website designs and metadata. Websites that utilize too many graphics and little indexable text create barriers for search engine crawlers much in the way they create barriers to visually disabled users. Crawlers have been characterized as "users with substantial constraints; they can't read text in images, can't interpret JavaScript or applets, and can't 'view' many other kinds of multimedia content" (Hagans 2005). It is worth noting that websites designed to address accessibility for disabled humans simultaneously solve problems of accessibility for machines.

The text that many digital repositories offer in the form of metadata can also be problematic, as the descriptive metadata used for one object is often indistinguishable from the next. This can be especially true in photograph collections where the same caption is used for numerous photos. Metadata field definitions are also often applied inconsistently, creating difficulty for search engines that must normalize data. In addition to inconsistent application of metadata fields, repository managers sometimes even use metadata schema that are not recognized or desired by search engines. Google Scholar may be considered the premier search engine on the open Web for academic and scholarly publications (Khabsa and Giles 2014), but it has difficulty harvesting content from many institutional repositories (IR) because of library metadata practices that relied on the Dublin Core schema. Google Scholar requires individual fields for each part of a citation and the Dublin Core schema simply doesn't provide those fields. As a result, Google Scholar recommends using other schema, such as Highwire Press, PRISM, Eprints and BePress (Arlitsch and O'Brien 2012).

Further technical SEO issues that may prevent digital repositories from being included in search engine indexes can be characterized as the "user experience." Search engine users are customers of search engine companies until the companies feel confident in directing them to specific websites. Websites that deliver a poor user experience in terms of slow server speed, dead links, or unexpected results may be excluded from search engine indexes or suffer from low rankings in SERP (Singhal and Cutts 2010; Brutlag 2009).

While Google is not the only search engine on the Web it has been the most dominant for more than a decade in North America and Europe. Fully two thirds of all search engine queries in North America are registered with Google properties. Yahoo! and Bing combine for nearly thirty percent and Ask Network and AOL, Inc. comprise approximately 4% (comScore, Inc. 2016). Google's market share in Europe is estimated at a startling 90% (Meyer 2015), cementing its status as the dominant search engine on both continents. Google is also the first of the major search engines to have leveraged the power of the Semantic Web through a knowledge graph and therefore is the primary search engine for which website and repository content should be optimized. However, Bing is close behind and others will follow; and in all likelihood they will tap similar data sources that currently help to populate the knowledge graphs that Google and Microsoft are developing.

Section 2.4 Semantic Web Optimization

The Semantic Web is a relatively recent development in the evolution of the World Wide Web that was launched in the early 1990s (Berners-Lee et al. 1994), and it has long promised a richer user experience than previous iterations of the Web. The first generation of the Web that was introduced to the world in the early 1990s was essentially a read-only format and has since become known as Web 1.0. Its second iteration (Web 2.0), is now referred to as the "read-write" Web and featured the first publishing platforms, like blogs and wikis, as well as development tools like JavaScript and XML. (Aghaei, Nematbakhsh, and Farsani 2012). Web 3.0 is the Semantic Web, and it offers the potential for more accurate, relevant, and enhanced search results, as well as syndication of data to numerous semantically-oriented applications and websites. "While search engines have long connected people to *documents*, they are now beginning to also connect people directly to *information*" (Bernstein et al. 2012).

Some of the first discussions about the necessity and capability of semantic data on the Web came from its acknowledged founder and visionary, Sir Tim Berners-Lee: "useful data on the Web would have to be available in a machine-readable form with defined semantics…" (Berners-Lee 1996). The promise of improved machine comprehension depends on "sufficient context about resources on the Web" that help machines "find the right things and make decisions" (Matthews 2005). In other words, the Semantic Web sets the stage for machine learning, and for this to happen "computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning" (Berners-Lee, Hendler, and Lassila 2001). Since Berners-Lee began explaining his vision in the mid 1990's, the Semantic Web has been the topic of much theoretical discussion, but in the past five years it has enjoyed dramatic development.

Whereas the previous generations of the Web have operated in a "strings-based" environment, the Semantic Web is commonly understood to be an "entity-based" environment in which machines interact with data records to better understand concepts, identities, and relationships of entities. These data records are expressed through linked data that is intended "to enrich the Web with structured data" (Thalhammer and Rettinger 2014). The metaphorical phrase "strings to things" alludes to the transition to the Semantic Web, in which search engines must evolve from matching user search queries to textual strings by algorithmic means, to a new process in which queries are matched with entity identification and entity relationships. "Today major search engines…want to understand the user queries semantically and serve their information needs precisely from their knowledge repositories" (Uyar and Aliyu 2015) and "Search engines no longer only return documents — they now aim to return direct answers" (Vaish et al. 2014).

As defined by Berners-Lee the Semantic Web requires de-referenceable addresses so that every established fact can become a destination that will help machine learning. "On the Semantic Web, all information has to be expressed as statements about resources… identified by Uniform Resource Identifiers (URIs)" (Sauermann, Cyganiak, and Völkel 2011). In this way the resources can become recognized as "entities" with verifiable information about things with a "distinct and independent existence" (Meij, Balog, and Okijk 2014). Linked data is the means that can help machines understand entities and their relationships on the Semantic Web. It requires every subject or object to have a URI, to be reachable through HTTP, to include useful information about the subject or object and its relationships

using the Resource Description Framework (RDF), and to refer to other things using their HTTP URI names (Berners-Lee 2006). Semantic triples codify "statement[s] about semantic data in the form of subject-predicate-object expressions" ("Semantic Triple" 2016). DBpedia is an example of a structured data knowledge base that offers its records as linked data in the form of RDF. Machines can understand information represented as linked data far better than they can understand information in unstructured text.

Changing machine comprehension of information on the Web from interpretations of strings of text to recognition of established entities that have relationships to other entities is crucial for accurate and robust representation of academic organizations. "These entities are not documents on the web, but rather constructed information about real world objects and concepts...The relationships of entities are particularly important" (Uyar and Aliyu 2015).

Graph databases can help traditional search engines move into the semantic search engine market by creating a place where facts about the entities and their relationships can be stored and tapped. The large databases being developed by search engine companies to gather information about entities are built on graph database models that are best equipped to manage the large volume of data that is beyond the scope of relational databases (Ali and Padma 2016). Graph databases are "applied in areas where information about data interconnectivity or topology is more important, or as important, as the data itself" (Angles and Gutierrez 2008). The knowledge graphs generated from these databases have been characterized as the "backbone of semantic search" (Meij, Balog, and Okijk 2014). The race to develop semantic search engines can be seen in the major commercial players, as "interest has been strongly growing, with evidence by projects like the Google Knowledge Graph, the EntityCube/Renlifang project at Microsoft Research, and the use of public knowledge bases for type coercion in IBM's Watson project" (F. Suchanek and Weikum 2013).

Google and Bing both launched knowledge graphs in 2012 as a way to gather and construct "information about real world objects and concepts" (Uyar and Aliyu 2015). Google's Knowledge Graph is populated with semantically rich data so that it can provide users with more accurate and enhanced search results (Singhal 2012). The Knowledge Graph itself is not visible to the public, but to date Google has produced three Knowledge

Graph-generated products that are visible in SERP: Answer Box; Carousel; Knowledge Graph Card. Examples of the first two are shown below.

The Answer Box provides definitions to concepts and displays at the top of organic search results (Perrott 2015) (see Figure 3).

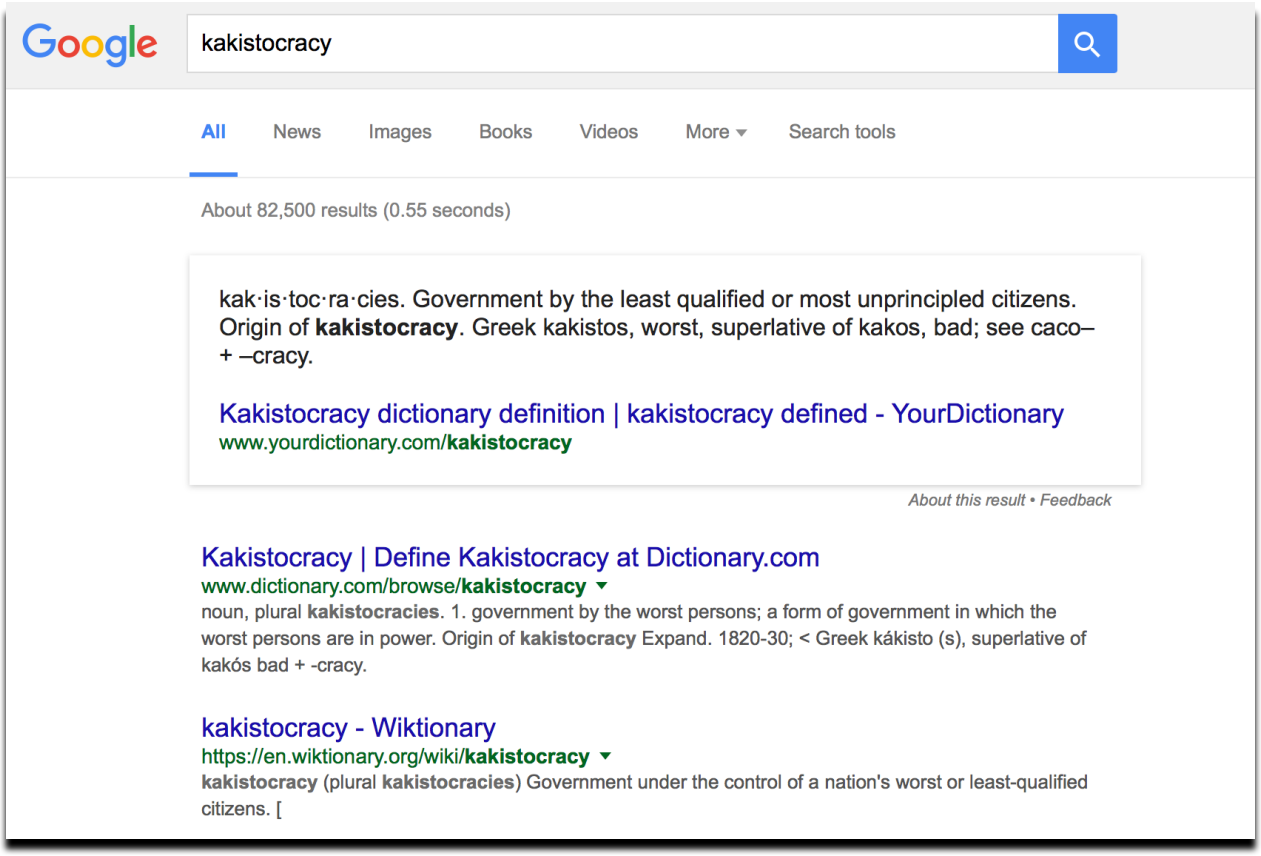

Figure 3: Google SERP showing an answer box above search results

The Carousel displays a set of instances that comprise an entity and is displayed across the top of the screen in SERP (Gesenhues 2013). In the example below, a search for "U.S. research universities" offers a carousel display of major research universities in the United States above the traditional list of search results links (see Figure 4).

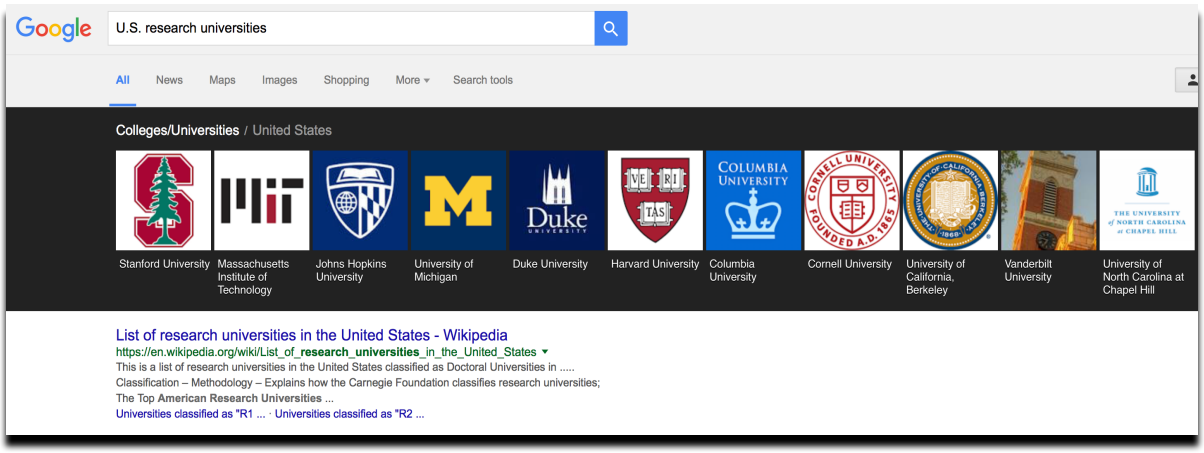

Figure 4: Google SERP showing a carousel display above search results

Both visible products of the Knowledge Graph help to enhance SERP by providing facts about entities that the Google search engine has determined are related to user search queries. The research in this dissertation focuses on the third visible product of the Knowledge Graph by demonstrating that KC (see Figure 2) do not consistently appear for academic organizations, and that KC are sometimes populated with inaccurate facts. The sources from which knowledge graphs draw their information to populate KC are examined in the next section.

Section 2.5 Sources of Information for Knowledge Graphs

The knowledge graphs that support search engines may have trouble confirming entities and their relationships without the help of "the publication of interlinked datasets on the Web, in a form that enables people and computer programs to use these datasets for navigation, integration, and web-scale reasoning" (Bouquet, Stoermer, and Vignolo 2012). Search engines seem most successful in gathering accurate information when organizations, people and concepts are defined and verified as entities in data sources the search engines trust. Over 1,000 knowledge bases currently comprise the LOD cloud, and many of these knowledge bases contain information about entities from which search engines can learn (Schmachtenberg, Bizer, and Paulheim 2014). The knowledge bases considered to be the most significant are represented near the center of the LOD cloud (see Figure 5), and include DBpedia (Auer et al. 2007), Wikipedia (Lih 2009), YAGO (F. M. Suchanek, Kasneci, and Weikum 2007), the CIA World Fact Book (Central Intelligence Agency 2015), Freebase

(Bollacker et al. 2008), and Wikidata (Erxleben et al. 2014). Google has acknowledged that its knowledge graph draws some of its information from some LOD knowledge bases (Sullivan 2012; Singhal 2012).

Wikipedia has developed into one of the most crucial sources for entity data on the Semantic Web. Within a few years of its launch in 2001 it had become the world's largest encyclopedia and was 25 times as large as Encyclopedia Britannica. It was also the sixth most popular website in the world (Messner and DiStaso 2013). But Semantic Web developers soon realized that its wealth of information was not accessible or useful to machines because it was built for human readability. "*Using* Wikipedia currently means *reading* articles - there is no way to automatically gather information scattered across multiple articles…its meaning is unclear to the computer, because it is not represented in a machine-processable, i.e. formalised way" (Völkel et al. 2006). A LOD knowledge base that contained machine-readable structured data records generated from the wealth of information in Wikipedia was required. Völkel and his co-authors proposed a "semantic Wikipedia" to help fill this gap and they called it DBpedia.

DBpedia was launched in 2006 and currently contains over 1.8 billion facts extracted from Wikipedia that are made widely available as structured data "via established Semantic Web standards and Linked Data best practices" (Lehmann et al. 2015). In practice, much of the information is extracted from Wikipedia infoboxes, which "display an article's most relevant facts as a table of attribute-value pairs on the top right-hand side of the Wikipedia page" (Bizer et al. 2009). DBpedia developers have manually created an ontology based on the most common infobox elements in Wikipedia ("DBpedia" 2013).

DBpedia has contributed enormously to the development of the Semantic Web by creating a structured data source from Wikipedia, without involvement from Wikipedia content creators. "The DBpedia project showed that a rich corpus of diverse knowledge can be obtained from the large scale collaboration of end-users, who are not even aware that they contribute to a structured knowledge base" (Bizer et al. 2009). Other knowledge bases in the LOD cloud, such as YAGO (Yet Another Great Ontology), also make use of Wikipedia and acknowledge the value of its infoboxes and category pages, although the YAGO approach differs from that of DBpedia: "…rather than using natural language processing on the articles of Wikipedia, our approach builds on Wikipedia's *infoboxes* and *category pages*" (F. M. Suchanek, Kasneci, and Weikum 2008).

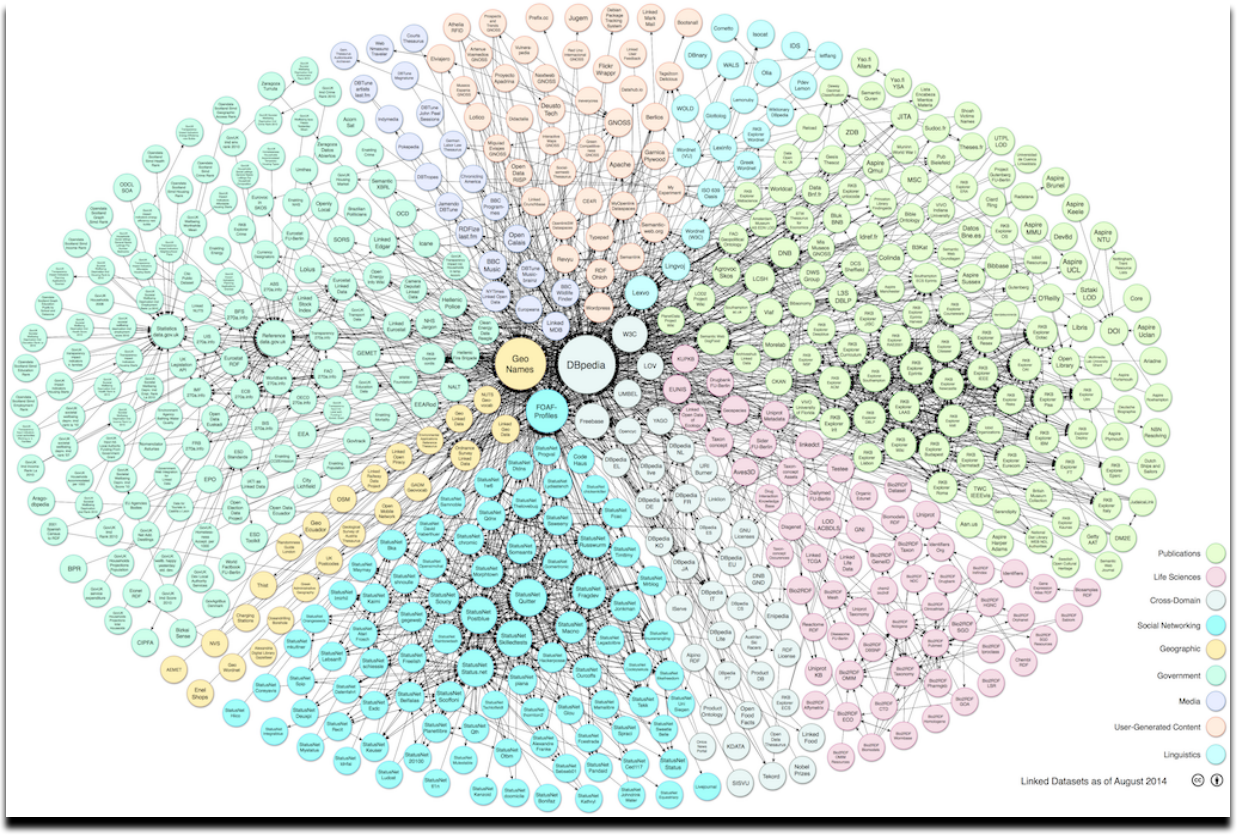

Figure 5: Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

Freebase was another massive data source that, until recently, was "one of the most popular knowledge bases (as evident by its use by major commercial search engines such as Google and Bing)" (Tan et al. 2014). Google acquired Freebase in 2010 when it purchased MetaWeb (Bergman 2012), but in 2014 Google announced that it would cease building Freebase in favor of Wikidata, a new community-supported knowledge base. A mass migration of Freebase records to Wikidata was completed in 2016 (Tanon et al. 2016a). At the time of this writing Freebase itself is still available, but in "read-only" mode.

Wikidata is a sister project to Wikipedia and is fast becoming a significant source of structured data records. Whereas DBpedia extracts structured data from Wikipedia to generate a linked data ontology that can help populate knowledge graphs, Wikidata does not directly extract data from Wikipedia. Instead, it is an open knowledge base that anyone can edit (Morrison 2013). Launched by the Wikimedia Foundation in October 2012, it uses a crowdsourcing model to create and edit structured data records while reconciling data from various Wikipedia language versions. It has over 3,500 active contributors who make a half

million edits per day, and the data are exposed in machine readable formats such as JSON, XML and RDF (Vrandečić and Krötzsch 2014). Wikidata administrators acknowledge its role in helping to populate Google's Knowledge Graph, but they are quick to point out that it does not replace Freebase. "Whereas Freebase was the open core of the Knowledge Graph, this is not true for Wikidata. Wikidata is one source of the Knowledge Graph among many, but does not have the same standing as Freebase had" (Wikidata 2015). Some SWO consultants maintain that Wikidata does influence KC results, but stress that there are no guarantees (Edward 2015).

Structured data knowledge bases in the LOD cloud have been crucial to the development of the Semantic Web, but it is difficult to know exactly which sources Google and Bing tap for their respective knowledge graphs. For instance, there is no evidence from Google that it draws directly from DBpedia. While acknowledging that its Knowledge Graph uses "public sources such as Freebase, Wikipedia and the CIA World Factbook (Singhal 2012), Google says only that its Knowledge Graph is connected to DBpedia by "transivity," meaning that the relationship is indirect and established only insofar as it draws from sources that do have direct relationships with DBpedia (Mendes and Jakob 2012). Microsoft also only hints at its use of LOD knowledge bases, such as this quote from its patent application for a process it developed for entity detection and disambiguation: "…the entity-based search system recognizes particular content sources as authoritative sources for discovering entity information. For example, the system may identify Wikipedia as having particularly strong and trustworthy entity information and may recognize various pages at that site as describing entities" (Li et al. 2013).

There is recognition in the Semantic Web community that problems of accuracy can occur with crowdsourced knowledge bases, which may cause search engines to be cautious in their utilization of these sources. For instance, "usage" of infoboxes within Wikipedia has been characterized by some as "chaotic" (Jentzsch 2009), meaning that article authors and editors create and populate infoboxes inconsistently, and often fail to take advantage of available templates. Some researchers have simultaneously confirmed the use of these knowledge bases by search engines while also warning of their shortcomings: "Such errors can have significant consequences because these knowledge repositories … often serve as data sources for third-party applications, such as Google's Knowledge Graph, Bing's Satori, and Facebook's Entity Graph" (Tan et al. 2014). Developers also recognize that much more

needs to be done to provide data for the Semantic Web, particularly at the level of websites themselves: "…despite massive undertakings such as the Google Knowledge Graph, Bing Satori and Freebase, much of the knowledge on the web remains unstructured and unavailable for interactive applications... While crowdsourced undertakings such as Freebase and DBpedia have created large amounts of structured data, they tend to only acquire high-level information and do not have enough contributors to achieve significant depth on any single entity" (Vaish et al. 2014).

The paradigm shift from traditional Web search that relied almost entirely on keywords, to semantic search that relies on knowledge graphs has affected the practice of traditional SEO. In addition to harvesting facts from knowledge bases, search engines are seeking richer descriptions and more authoritative comprehension of entities from the sites they harvest, and in fact are beginning to favor sites that deliver descriptions that are comprehensible to machines. "To optimize websites for search in the future, SEOs will need to create relevant, machine-recognizable 'entities' on webpages that answer well-refined, focused or narrow queries" (Bruehmmer 2013). Website developers are responding by implementing recommended schema and syndication methods, most notably through Schema.org markup offered through JSON-LD or RDFa syndication. Thus, "Semantic Web Optimization" is the practice that can help organizations become relevant to search engines in this new environment, and covers both the creation and monitoring of LOD knowledge base records and the on-page markup required in websites (Lloyd 2014).

Section 2.6 Knowledge Graph Cards

Knowledge Graph Cards (KC), also known as Knowledge Panels and Knowledge Cards, have seen increasing discussion in the blogs and websites of search engine optimization firms and consultants since Google's introduction of the Knowledge Graph in 2012. However, little information about them or how they are generated can be found in the scholarly literature.

Google defines the KC as "an information card [that] appear[s] along with regular search results" and "can show up for a wide range of topics, including certain companies, products, celebrities, groups, movies, and TV shows" (Google, Inc. 2016c). Specific directions on influencing the display of KC are still minimal, but in 2016 Google began offering a "Suggest an edit" link on KC and additional instructions on its support site (Google, Inc.

2016c). However, these instructions are only useful when a KC already exists, and one consultant notes that "some parts of the Knowledge Graph Card can't be edited, like the subtitle, Wikipedia snippet, images other than the main image, and 'People also search for'" (Edward 2016). Google has a patent that generally describes a process whereby a user can update a "personal knowledge panel" and the "system allows the user to update information directly in the knowledge panel interface without the need to visit another web page" (Trew, Swerdlov, and Lai 2016). A few researchers have tried to reverse engineer the KC, some of it mere speculation (Bergman 2012) and some applying more sophisticated analysis methods (Assaf et al. 2014). In general, no studies assessing the appearance of KC for groups of entities, such as academic libraries or broader academic organizations, are evident in the scholarly literature.

Section 2.7 Semantic Web Identity

The term "Semantic Web Identity" does not yet appear in the scholarly literature except where the author and his colleagues introduced it in its earlier form of "Semantic Identity" (Arlitsch et al. 2014). The author characterizes Semantic Web Identity (SWI) as the condition where search engines have recognized the existence and the interests of an entity that has a presence on the Web. An entity can be a person, organization or concept, but for the purposes of this research it is defined as an academic organization in the hierarchy of the parent institution, i.e., a college, department, center, or institute.

"Digital identity" is a related term that appears in the literature, but it is mainly used to describe reputation management (Izenstark 2014). Professional organizations sometimes encourage their members to establish and actively manage their digital identities in response to third party sites that independently review, rank, or describe the members. Techniques advocated for physicians, for instance, include creating their own websites, populating professional media services like LinkedIn, or engaging in social media applications like Twitter and YouTube (Gill, Zampini, and Mehta 2015). While not directly related to the research described in this dissertation, the concept of engaging with structured data records in knowledge bases is similar.

Another related term is "researcher profile," which is more specific than the engagement with professional and social media services described above and is useful for

researchers who want to draw more attention to the published results of their work. The first intent of researcher profiles is to disambiguate the researcher from colleagues with similar names (Cals and Kotz 2008), usually through the use of identifiers in established systems such as ResearcherID (Enserink 2009) or ORCID (Haak et al. 2012). The second intent is to gather the researcher's publications in a publicly accessible service, such as Google Scholar, ResearchGate, Academia.edu, or open access institutional repositories. Ideally, the creation of a profile in these services triggers an automated harvesting process of the researcher's publications, thereby minimizing laborious manual data entry (Shanks and Arlitsch 2016). Some of these repositories support the creation of robust profiles through biographies or curriculum vitae that may find their way into search engine knowledge graphs for those individuals.

While digital identity and researcher profiles are related issues, they are not directly tied to this research regarding SWI for academic organizations.

Section 2.8 Action Research Methodology

Western philosophical traditions describe empirical knowledge as based on verifiable observation, and empiricists go so far as to maintain that "all human knowledge is derived from experience" (Duignan 2015). Results of research for this dissertation can be empirically validated through a multi-methodology known as "action research," which originated in the field of social sciences but gained adoption in other academic disciplines, including information systems (Baskerville 1999), management, and library and information science. Kurt Lewin first introduced action research as a bridge between social theory and action: "a comparative research on the conditions and effects of various forms of social action, and research leading to social action" (Lewin 1946). It was initially applied by Lewin (and later, Lewin and the *Tavistock Institute*) to study psychological and social disorders among veterans of battlefields and prisoner-of-war camps (Baskerville 1999).

Action research in its social sciences domain has been defined as "a spectrum of activities that focus on research, planning, theorizing, learning and development" (Cunningham 1993). Some characterize it rather succinctly as "a cyclical inquiry process that involves diagnosing a problem situation, planning action steps, and implementing and evaluating outcomes (Elden and Chisholm 1993). Stringer positions the methodology as

pragmatic, setting the stage for its use in information systems and library and information science disciplines: “Action research is solutions-oriented investigation leading to resolution of issues investigated” (Stringer 2014).

Sanford adds complexity to the cyclical and recursive nature of the methodology, explaining a process of “analysis, fact-finding, conceptualization, planning, execution, more fact-finding or evaluation; and then a repetition of this whole circle of activities; indeed, a spiral of such circles” (Sanford 1970). Lewin’s contention that “research that produces nothing but books will not suffice” (Lewin 1946) continues to resonate, reminding us that knowledge produced from research has little meaning without action. Indeed, some see it as the best of both worlds. “Action researchers have developed new ways of thinking about research that would solve practical problems and contribute to general scientific theory (Elden and Chisholm 1993). Dickens and Watkins acknowledge that while “action research consists of cycles of planning, acting, reflecting or evaluating, and then taking further action” it is also true that “the literature offers a variety of applications of action research.” Dickens and Watkins themselves write in the management literature to describe action research in the context of organizational development (Dickens and Watkins 1999). Even in its purely social sciences application, then, action research as a methodology supports a computer science/information science paradigm of iterative testing, re-evaluation and repositioning based on results. “Action research is also characterized by a commitment to effect real change, and an iterative approach to problem solving” (Easterbrook et al. 2008). Action research in the social sciences has also developed its own sub-methodologies, including participatory action research (PAR) and canonical action research (CAR) (Davison, Martinsons, and Kock 2004).

Section 2.8.1 Action Research in Information Systems

The academic discipline of Information Systems (IS) is closely related to Computer Science and Informatics, but it deals more specifically with “the use of information and communications technology in organizations” (Davis 2006). Action research is one of the most common nontraditional methodologies applied in the IS literature (Mingers 2003) and has been included in a taxonomy of IS research approaches (Galliers and Land 1987). Following a trace of the epistemology of action research to pragmatism, Nielsen proposes six criteria for designing and evaluating action research for information systems: roles,

documentation, control, usefulness, frameworks and transferability (Nielsen 2007). Other researchers (Baskerville and Wood-Harper 1996) have enumerated three characteristics of the ideal domain of the action research method for the IS discipline:

1. The researcher is actively involved, with expected benefit for both researcher and organization;
2. The knowledge obtained can be immediately applied;
3. The research is a cyclical process linking theory and practice.

These steps apply to the research conducted for this dissertation, as the knowledge that was acquired by the author/researcher was concurrently applied in several case studies that aimed to improve the SWI of the subject organizations. The SWI improvement process was modified as more knowledge was gleaned from data collection and application of techniques, revealing a cyclical process that was successful in improving the appearance of the KC.

Section 2.8.2 Action Research in LIS

The scholarly literature also provides evidence that action research is useful for library and information science (LIS). LIS is multi-disciplinary, and other methodologies that support the study of information users and their interactions with technologies have sometimes been deemed "unsatisfactory" (Wilson 2000). The flexible and multi-method approach of action research may be applied to both qualitative and quantitative evaluation, and therefore is ideal for LIS. "Action research…is a type of research that focuses on questions or problems in the workplace and attempts to find answers that solve, or shed light on, the specific problem under study" (Farmer 2011). Adapting the action research methodology to LIS, Cook offers a checklist, while cautioning that not every item in the list will be used for each study (Cook 2011)

1. Focus on an issue
2. Review theory
3. Develop questions
4. Collect data
5. Analyze data
6. Report results
7. Design action plan
8. Take action
9. Evaluate action

Wilson states that “changes in technology may enable an information service to perform tasks in service for the user not possible previously” (Wilson 2000). Connaway and Powell speak to the potential of data derived from action research being used to “improve a service, develop a new one, or discontinue a service” (Connaway and Powell 2010). These statements directly support a research goal documented in this dissertation, which encourages libraries to consider the possibility of offering new SWI services other academic organizations.

The action research methodology proposes specific actions (steps 7-9 in Cook’s list, above) to address the research findings and to make a real-world impact. Although case studies will be described in this dissertation to support the hypotheses, along with a description of a new SWI service that is being implemented at MSU, the emphasis of this work is scholarly and it is not intended to prescribe policy or specific practices. Instead, the case studies and description of the service are offered merely to demonstrate what could be achieved if libraries agree that achieving SWI for their organizations and their campus constituents is desirable. Specific actions should be designed, implemented, and evaluated locally and in further study.

Section 2.9 Summary of the Scholarly Context

This chapter evaluated the larger context in which the research is conducted by reviewing the scholarly and non-scholarly literature in several related areas. The review helped to establish that this work delves into new areas of scholarship and combines aspects of LIS, business marketing, and computer science. The aspect of marketing that is most related to establishing SWI is that which deals with consistent use of organization names and branding. The LIS literature fails to adequately discuss this issue, particularly with regard to the Semantic Web.

SEO is the practice of techniques that improve the visibility of websites and digital objects in SERP. It can be considered the foundation of the research in this dissertation, although the environment of the Semantic Web requires a modified approach that has become known as Semantic Web Optimization (SWO). Semantic search engines increasingly rely on graph databases to assemble facts about entities, and Google’s Knowledge Graph and Bing Satori are two examples of major search engine graph databases. Google has

produced three visible products from its Knowledge Graph: Answer Box; Carousel; and Knowledge Graph Cards (KC). Examples were shown of each of these products, but this research focuses on the KC. The scholarly literature concerning the sources that influence KC is limited, but there is enough to establish hypotheses that were described in Section 1.3.

Action research originated in the social sciences as a research methodology intended to catalyze social action from theory. It has since found use as a research methodology for other disciplines, including information systems, management, and library and information science. Its flexibility and cyclical approach to solving workplace problems that require empirical evaluation of quantitative and qualitative data make it an ideal methodology for verifying techniques to establish and improve Semantic Web Identity for academic organizations.

An argument can be made that this research focuses too much on the products of a company whose dominance of the search engine market may eventually wane. Many search engine companies have come and gone, or have at least seen their products wither ("Timeline of Web Search Engines" 2016), and even a technology giant like Microsoft Corp. has found itself at the short end of the search engine market share, a fate no one would have predicted twenty years ago. Microsoft's Bing search engine is addressing similar Semantic Web search problems, but as of this writing Microsoft still has not achieved the search engine market dominance of its rival, Google. Future studies comparing SWI of academic organizations in Google and Bing would be useful, but for the moment that effort is beyond the scope of this work.

The Internet as we currently know it would look very different without the savvy and inventiveness of Google, which has held the dominant position in the search engine market for fifteen years and seems well positioned to continue that hold. Longevity is a rare quality in the competitive world of Internet-dependent businesses, where even the proverbial ground those businesses are built on is constantly shifting and evolving due to new developments that are improving networking, processing, and storage technologies. It would be foolish to try to predict how long Google's Knowledge Graph will be the cutting-edge technology that helps it draw the most users to its search engine properties, but it is equally foolish to watch these fantastic developments unfold without learning how to leverage them now for the competitive benefit of libraries and other academic

organizations. Investments in Semantic Web data sources are also likely to result in other benefits in the many applications that are being built to utilize those sources.

The next chapter will describe the specific methods that will be applied to the study under the guidance of the action research methodology.

Chapter 3 Research Methods

Section 3.1 Introduction

This chapter describes the methods that were used to collect and analyze data in response to the research questions. The process uses elements of reverse engineering to reveal some of the data sources that Google uses to generate and populate Knowledge Graph Cards (KC). Guiding the methods of this research are the nine steps that Cook used to adapt the action research methodology to the LIS discipline, which were listed in the previous chapter (Cook 2011). Each of the nine steps is listed again in Figure 6, and the sections in this chapter will be aligned with those steps, showing how they are applied to the current research.

Section 3.2 Action Research Design

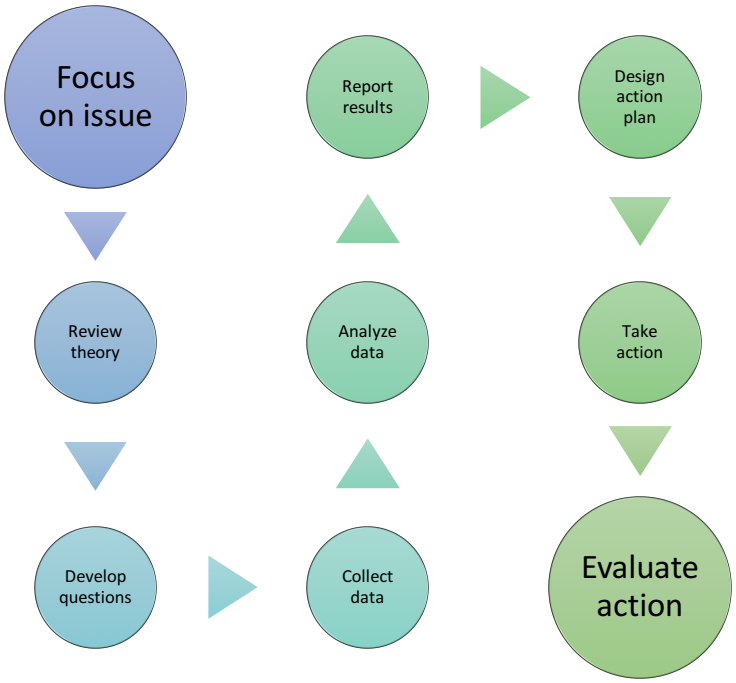

Figure 6: Action Research Methodology

Section 3.2.1 Focus on an Issue

SWI may be said to exist when the search engine has established enough verified facts

about an organization to display them in the form of a KC. The existence of SWI is not absolute, but may be perceived on a sliding scale of robustness in the number of verified information elements (referred to in the statistical analysis as "outcome variables") that are displayed in the KC. Generally, the more robust the KC, the more the organization can be said to have achieved SWI. This dissertation hypothesizes that populating certain LOD and proprietary knowledge bases will compel the generation of a KC if one does not currently exist. If a KC currently exists, populating knowledge bases is expected to enhance the robustness of the KC.

This research surveys the current condition of Semantic Web Identity (SWI) for member organizations of the Association of Research Libraries (ARL). It provides statistical evidence to help libraries decide where to focus their energies if they wish to establish or improve SWI for their organizations. It also provides some illustration that SWI concerns extend beyond libraries and into other academic organizations. Chapter 6 examines the broader implication of the research, including a description of an SWI service developed at MSU that libraries may consider implementing locally to help academic organizations on their campuses achieve SWI.

The ARL was chosen because it is the premier membership organization for research libraries in North America. Founded in 1932 "exclusively for literary, educational and scientific purposes by strengthening research libraries" (George and Blixrud 2002), the ARL currently consists of 125 research libraries (Association of Research Libraries 2016) whose invited membership is "distinguished by the breadth and quality of their collections and services" and "the parent institution's aspirations and achievements as a research institution" (ARL Board 2013). The author's home library at Montana State University is not a member of the ARL, largely because its personnel expenditures do not meet the membership threshold.

Section 3.2.2 Review Theory

It is anticipated that some Linked Open Data (LOD) and proprietary knowledge bases help populate graph databases managed by search engines. Graph databases - such as Google Knowledge Graph and Bing Satori - establish verified facts that are drawn (at least in part) from knowledge bases that help the search engine verify the existence of organizations and gain understanding of the nature of their businesses. LOD knowledge bases examined in this

study are Wikipedia, DBpedia and Wikidata, while the proprietary knowledge bases are Google My Business and Google+. Search engines may also take an organic approach to gathering facts by harvesting information from websites, but this approach may be more prone to error as few academic websites currently contain structured data in Schema.org.

Section 3.2.3 Develop Questions

The following research questions were introduced in Chapter 1, and are reproduced here to illustrate their fit into the action research methodology. The questions are designed to fulfill the research goals of demonstrating the current condition of SWI among ARL libraries, and to provide statistical evidence that may help libraries address their own SWI as well as the SWI of other organizations at their institutions.

Research Question 1: What is the current state of Semantic Web Identity of ARL libraries, as indicated by the presence of accurate Knowledge Graph Cards in Google search results when the primary and alternate names of those libraries are searched?

Research Question 2: Are records or profiles present for ARL primary and alternate library names in the following knowledge bases: Google My Business; Google+; Wikipedia; DBpedia; and Wikidata?

Sub-question 1: Is an accurate KC likely to display in search results if the library organization has not been claimed and verified in Google My Business?

Sub-question 2: Is a KC likely to display a description field (one possible information element) if no Wikipedia article exists for the primary or alternate name of the library?

Research Question 3: Does the presence of a given knowledge base record predict the odds of other facts (information elements) on the KC being populated?

Section 3.2.4 Collect Data

A data set was collected for the 125 members of the Association of Research Libraries (ARL). Collecting data was accomplished by conducting Google searches for the organizations, observing the presence and robustness of their KC, and capturing evidence of those results with screen capture software. Searches were conducted from the city of Bozeman, Montana, in the United States, using the main United States-based Google search engine[1].

[1] U.S. Google search engine – https://www.google.com

The ARL libraries were searched by both their primary name (as listed in the ARL directory) and an alternate name when one could be determined, for a total of 219 names. The primary name of a library is considered its official name, and is supplied to ARL for the membership directory by the member library (Baughman 2016). For instance, the ARL lists in its members' directory[2] *Yale University Library* as the official name of the library at Yale University and this is confirmed by the title of the website[3] at Yale University. However, most universities that host multiple libraries on their campuses have one that is considered the "main library," and often that library is known to the local community by a different name. At Yale University the local name is the *Sterling Memorial Library.* For the purposes of this study, the ARL listing (*Yale University Library*) was considered the primary name and the local name of the main library (*Sterling Memorial Library*) was considered the alternate name. Both names were searched and screen captures of the search results were collected for each. Screen capture image files of the search results included the search phrase, and the date of the search was captured in the filename in the ISO 8601 international date format: yyyy-mm-dd.

Five knowledge bases (Google My Business, Google+, Wikipedia, DBpedia, and Wikidata) were also searched for the 219 primary and alternate names of each organization, and the resulting records for each name in each knowledge base were again recorded using screen capture software. It is worth briefly reviewing the reasons for selection of these knowledge bases. Google My Business and Google+ were chosen because of Google's own explicit or implicit acknowledgement of their roles in helping Google realize and verify the existence and location of organizations. Wikipedia and DBpedia were selected because of their acknowledged prominence as LOD sources on the Semantic Web, and Wikidata was chosen due to its recent inheritance of Freebase records, since Freebase was a known data feed for Google's Knowledge Graph prior to its retirement. References can be found in the literature to other knowledge bases that may influence Google's Knowledge Graph (such as the CIA World Factbook), but the five that were chosen are easily accessible through public websites that can be queried to display records. The one exception is DBpedia, which offers sophisticated SPARQL query interfaces, but it does not have a common keyword search box

[2] ARL membership directory - http://www.arl.org/membership/list-of-arl-members

[3] Yale University Library – http://web.library.yale.edu

found on most websites. However, since DBpedia records are automatically generated from Wikipedia articles and inherit a similar URL structure, it is easy to reveal the DBpedia record by slightly altering the Wikipedia URL. An example follows:

> *Wikipedia article URL for Montana State Library*
> *https://en.wikipedia.org/wiki/Montana_State_University_Library*
>
> *DBpedia record URL for Montana State Library*
> *http://dbpedia.org/page/Montana_State_University_Library*

Among the data recorded in the spreadsheet were columns titled "AccurateKC" and "SameAs" The former indicated whether the KC that displayed was correct for the library being searched, while the latter was a score that indicated whether the same KC was displayed when both the primary and alternate name searches displayed a KC, indicating the desired semantic "same as" comprehension by the search engine. For example, a search for "Auburn University Libraries" displayed the same KC as the search for "Ralph Brown Draughon Library," indicating that Google understood that the primary and alternate names represented the same organization. The "same as" comprehension is an important marketing technique in the age of machine readable data records. While humans are capable of making mental associations to establish relationships, machines must be explicitly informed that those relationships exist. Libraries should be aware that there are mechanisms in some of the knowledge bases that facilitate establishing the "same as" relationship between their primary and alternate names.

Section 3.2.4.1 Robustness Scores.

Knowledge Graph Cards for organizations generally have similar information elements, but may differ based on the type of entity to which they are keyed (Slawski 2015). While no exact list of elements seems to exist for academic entities, some SEO consultants have compiled lists for businesses through observation, and these may include the following: basic info, photos, stock prices, reviews, social profiles, competitors and related searches (Dame 2015). While collecting data for this dissertation the author similarly observed certain recurring elements (see Figure 7) and observed that the information elements tend to appear in groups. It is rare for a single

element in a certain group to appear without the others, instead, either all the elements for a certain group appear or none appear. Because of this observation, the author identified eight common information elements for academic libraries and has organized them into three logical groups (see

Table 1). In statistical terms, the information elements are known as dependent or outcome variables, and they appear because of the presence or action of independent or explanatory variables. Research Question 3 asks whether the presence of any of the knowledge base records (i.e., independent variables) can affect the odds of observing the presence of the information elements (i.e., dependent variables) in the KC. The grouping of the information elements facilitates a more statistically useful calculation than if the analysis were run against the presence of every individual information element.

Since KC are designed to display relevant facts about the organization and how to locate it, the author structured seven of the eight information elements into two groups: Appearance and Contact. The third group (Description) may be considered somewhat anomalous because it is a free-text information field that Google states explicitly on the KC is drawn from a single source: Wikipedia. Therefore, the Description was scored as both an information element and a group.

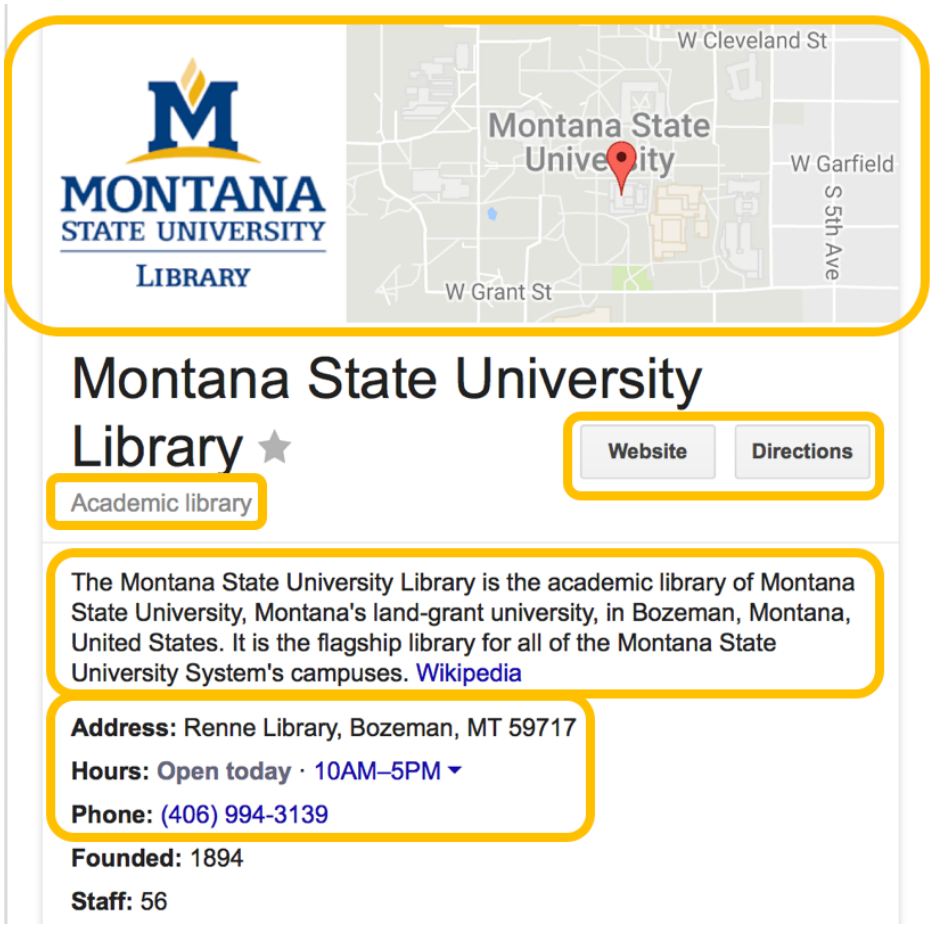

Figure 7: Sample KC showing most of the information elements that were recorded for ARL libraries

Table 1: Eight KC information elements categorized into three groups

Group	KC Information Element (Dependent Variable)
Description	Description
Appearance	Logo
Appearance	Image
Appearance	Type
Contact	Address
Contact	Telephone Number
Contact	Website link
Contact	Directions link

Data for a few other information elements that sometimes appear on KC were also recorded. While these could be analyzed later, they were not used in this study because the

presence of these elements was much more inconsistent and because they were deemed less useful to the user. The discarded elements included social media icons, user reviews, related searches, hours, and a designation called “located in” which, when present, offered the name of the parent institution of the organization for which the KC was displayed.

The presence or lack of the eight information elements in each KC was recorded and each group (Appearance, Contact, Description) was given a binary score by calculating the product of the information elements. A zero in any information element resulted in that entire group being scored as zero, as this helped strengthen the predicted odds of a given knowledge base having a positive effect on the group. The reason for binary scoring of the groups is that the presence of the information elements is binary - they are either present or they are not - and this dichotomous nature must be carried forward into the logistic regression model that predicts the odds that those groups affect the display of a KC. If the information elements could be scored along a scale, then a linear regression model would have been used to predict the robustness of the KC.

Section 3.2.4.2 Scoring principles for the records

The presence of records for the primary and alternate names of the ARL libraries in the five proprietary and LOD knowledge bases were also noted on the spreadsheet. The following list describes the scoring principles that were applied for each knowledge base:

- Google My Business – 0 indicates the business has not been claimed; 1 indicates the business has been claimed.
- Google+ - 0 indicates the lack of a profile; 1 indicates the presence of a profile; 2 indicates a verified profile, meaning it displays a checkmark added by Google (Saha 2013).
- Wikipedia – 0 indicates lack of an article; 1 indicates the presence of an article; 2 indicates an article that includes an infobox.
- DBpedia – 0 indicates lack of a record; 1 indicates presence of a record.
- Wikidata – 0 indicates lack of a record or a record that only includes a pointer to a Wikipedia article; 1 indicates a record with at least two populated fields.

Section 3.2.4.3 Other Scoring Principles:

A search for a library's primary name sometimes displayed a KC for its alternate name. In that case, a KC was said to exist (value=1 in the KC column) but it was not considered the

correct KC (value=0 in the AccurateKC column). For example: "Duke University Libraries" displays a KC for the "William R. Perkins Library," indicating the correct institution, but not for the specific search of the official name listed in the ARL membership directory. This phenomenon occured often and will be examined further in the Discussion chapter. For now, it is sufficient to say that ARL libraries represent themselves inconsistently in different fora, contributing to search engine confusion about their organizations.

The displayed KC was scored for robustness even if the primary name search displayed the KC for the alternate library name, or vice versa. However, if the search retrieved a KC that reflected neither the primary or alternate name then the KC was considered non-existent and was not be scored for robustness. For example: a search for "Boston University Libraries" displayed a KC for the "Music Library Boston University." This was clearly a branch library on campus that was neither the primary or alternate name of the organization being searched, and the KC was therefore considered non-existent.

Libraries that did not have their own Wikipedia articles, but rather were included as paragraphs or sections of their parent institutions' articles, were scored as having no Wikipedia article.

Many Wikidata records have apparently been automatically generated from Wikipedia articles but are unpopulated, containing only a reference back to Wikipedia. If the Wikidata record contained only this reference to Wikipedia then it was considered to be non-existent.

Section 3.2.4.4 Collecting Data for Other Organizations

The author has conducted spot checks with a number of universities, confirming that the condition demonstrated with ARL libraries exists across academic organizations, nationally, but a systematic collection of data to establish baseline measurements for a much larger grouping of non-library academic organizations is beyond the scope of this dissertation. Instead, a more cursory data set was collected from colleges at MSU to illustrate that lack of SWI extends into other organizations within academic institutions. MSU is representative of research universities in the U.S., and the data set provides a baseline measurement from which improvements in the SWI of MSU colleges can be shown relative to the service that the MSU Library has been offering to campus organizations for more than one year.

The MSU data set only included screen captures of the results of Google searches for the name of the organization, showing the presence or lack of KC. While the five knowledge bases were searched to determine whether data records exist for the MSU organizations, no scoring was conducted to measure the presence of the information elements. The eleven colleges that were surveyed represent the next organizational level below the overall institution, which is typical of the hierarchy of most research universities in the United States. Each college comprises multiple departments, and below departments there exist centers and research institutes. These are the eleven colleges that were surveyed:

- College of Engineering
- College of Letters and Science
- College of Arts and Architecture
- College of Nursing
- College of Education, Health and Human Development
- College of Agriculture
- Jake Jabs College of Business and Entrepreneurship
- Gallatin College
- Graduate School
- Library
- Honors College

Section 3.2.4.5 Software tools for collecting data

Data were collected in the Apple OS X (El Capitan) environment. The Google Chrome (v51, 64-bit) web browser was used to conduct searches for the presence of the KC in Google search results and for records from the LOD data sources: Wikipedia, DBpedia, and Wikidata. The browser's "Incognito" feature was turned on, which prevents Chrome from saving sites that have been visited and thus reduces the potential of search results being customized based on previous searches (Google, Inc. 2016a).

The Safari (v9.1) web browser was used to search for the presence of records in Google My Business and Google+. It is not possible to search Google My Business without being signed into a Google Account, and the Google+ display in a Chrome browser in Incognito mode would have necessitated two screen captures for each Google+ profile

because the "About" section displays over the main page in Incognito mode.

Screen capture software (*Jing* (version 2.7.0) 2014) was used to take snapshots of the screen when conducting Google searches and visiting knowledge base sites. Files were saved in the Portable Network Graphics (PNG) format and occasionally as PDF.

A spreadsheet to record and quantify the data collected was created with Microsoft Excel (v15.22). A simple binary measure indicating the presence (1) or the lack (0) of most records was recorded. Google+ profiles and Wikipedia articles received an additional score (2) that noted a verified profile (Google+) or the additional presence of an infobox (structured data) feature in Wikipedia.

Notes were taken during data collection and analysis, first using Evernote (v6.6) and then Microsoft OneNote (v15.22).

Section 3.2.5 Analyze Data

Data were analyzed in the R software environment for statistical computing and graphics (Gentleman and Ihaka 2016) using RStudio, an integrated and graphical development environment for R (*RStudio* (version 0.99.893) 2016). The Excel spreadsheet was converted to a CSV (comma separated values) file format for ingest into R.

Data quality was improved by running subset commands in R (Dalgaard 2002) to reveal inconsistencies in spreadsheet data values. These inconsistencies were then addressed individually by making second evaluations of the screen capture files to confirm or alter scoring in the spreadsheet. For example, the following subset command in R: `[subset(SWI,KC=="0"&AccurateKC=="1")` revealed library organizations for which no KC had been deemed to exist (value=0) as the author searched for each library name, but he had recorded a score of 1 in the "AccurateKC" column, indicating the supposedly non-existent KC was correct for the organization. The subset command in R revealed these inconsistencies, at which point the screen capture images were reviewed again and the scores were aligned to agree, either by acknowledging that a KC did exist or by changing the "AccurateKC" score to zero. Similar subset commands were run to reveal inconsistencies in other data values.

The following descriptive statistics statements were developed for the first two research questions, and these statements guided the calculations created in R and are used

to display the results in the Findings chapter:

Research Question 1: What is the current state of Semantic Web Identity of ARL libraries, as indicated by the presence of accurate Knowledge Graph Cards in Google search results when the primary and alternate names of those libraries are searched?

Descriptive Statistics Statements

1) Number and percent of accurate KC found for either the primary or alternate names of ARL member libraries.
2) Number and percent of accurate KC for ARL libraries that displayed the same KC for both primary and alternate names.
3) Number and percent of primary library names that displayed accurate KC, and number and percent of alternate library names that displayed accurate KC.

Research Question 2: Are records or profiles present for ARL primary and alternate library names in the following knowledge bases: Google My Business, Google+, Wikipedia, DBpedia and Wikidata?

Sub-question 1: Is an accurate KC likely to display in search results if the library organization has not been claimed and verified in Google My Business?

Sub-question 2: Is a KC likely to display a description field (one possible information element) if no Wikipedia article exists for the primary or alternate name of the library?

Descriptive Statistics Statements

1) Number and percent of libraries that have "claimed and verified their businesses" in GMB
2) Number and percent of libraries that have unverified Google+ profiles
3) Number and percent of libraries that have verified Google+ profiles
4) Number and percent of libraries that have Wikipedia articles without infoboxes
5) Number and percent of libraries that have Wikipedia articles with infoboxes
6) Number and percent of libraries that have DBpedia records
7) Number and percent of libraries that have Wikidata records

An explanation of the method applied to answer RQ3 follows a restatement of the question, below.

Research Question 3: Does the presence of a given knowledge base record predict the odds of information elements on the KC being populated?

Method: Odds were calculated through logistic regression analysis, which evaluates the relationship of the three groups of information elements with the presence of other knowledge base records. "Regression analysis determines the nature of the relationship and enables us to make predictions from it (Rowntree 2004).

- Logistic regression was calculated for the odds of appearance of the three groups of information elements against the independent variables of GMB, Wikipedia and Wikidata.
- Since the author's experience shows that successfully claiming a business in GMB will auto-generate a verified Google+ profile, logistic regression was not calculated against Google+. This helped reduce the potential effect of multicollinearity, which will be explained further in Section 4.2.3.
- Since the literature indicates that a Wikipedia article must exist before a DBpedia record can be generated, logistic regression was not calculated against DBpedia. Again, this helped reduce the potential effect of multicollinearity.

Section 3.2.6 Report Results

Findings are reported in the Chapter 4. The full data set that supports the findings is archived in Montana State University's *ScholarWorks* open access institutional repository (Arlitsch 2016). The data set includes:

- Readme file describing the contents of the data set (see Appendix E)
- Zipped archives containing more than 1400 screen capture files
- Two CSV spreadsheets
- Two R source frame files, containing all commands that were run for data analysis

In addition to results of equations that are rendered in table format, a type of graphical display called a “table plot” is used in the Findings and Discussion chapters to provide visual interpretations of the statistical analyses. Table plots are generated by RStudio and they show all 219 rows of the spreadsheet for the columns of spreadsheet data that are being compared with one another. Figure 8 is reproduced below, simply as an example. In this table plot the left-hand column displays the ARL libraries’ primary and alternate names in blue and orange, respectively. That column is compared with the number of accurate KC that appeared for each of those names. The table plot gives an easy visual demonstration that more accurate KC (right-hand column, green rows) appeared for ARL alternate library names (left-hand column, orange rows) than for the primary names (left-hand column, light blue rows).

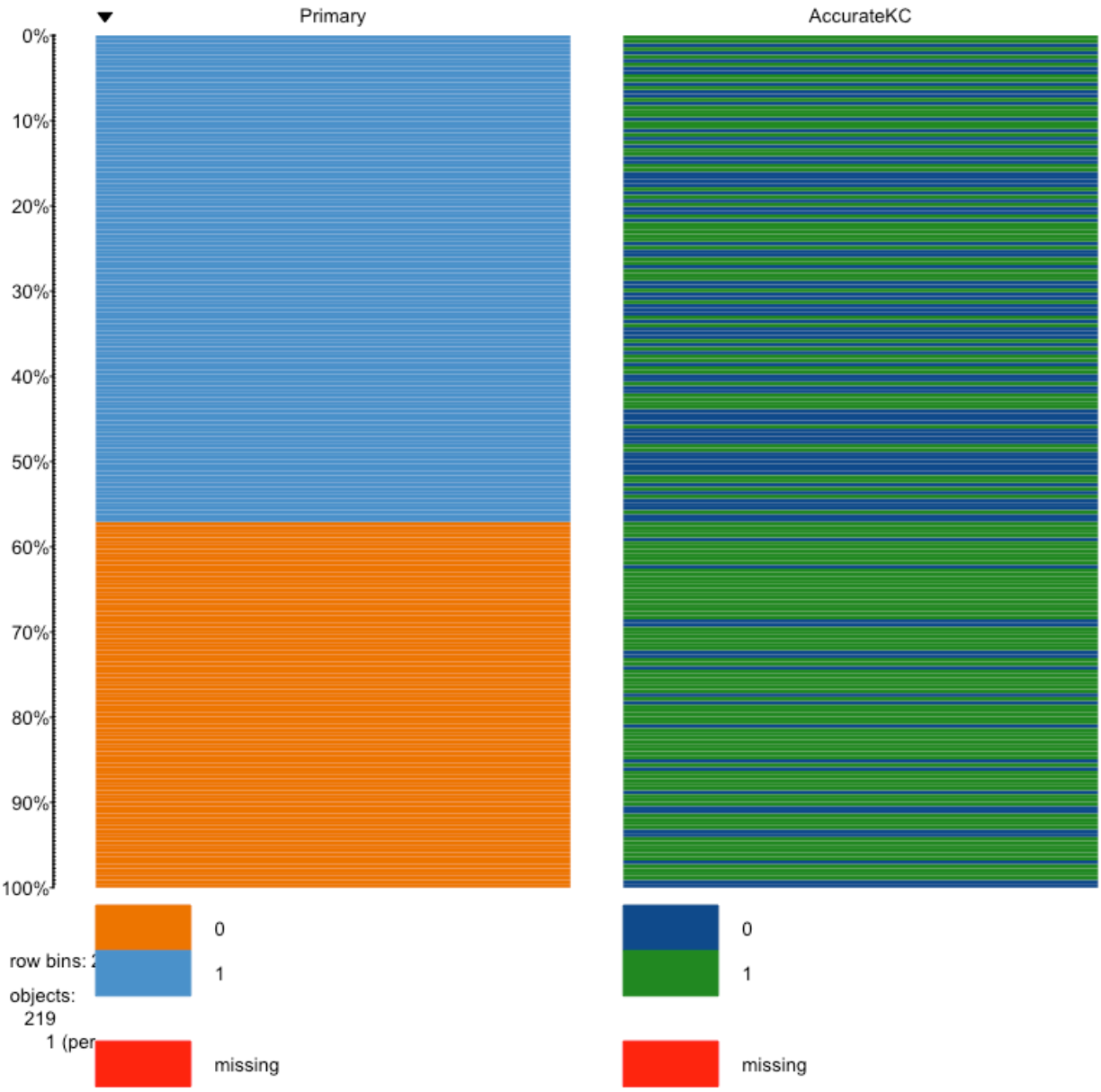

Figure 8: Table plot showing that ARL library alternate names (column 1, orange rows) were more likely to display an accurate KC (column 2, green rows)

Section 3.2.7 Design Action Plan

Action research methodology was developed to bridge theory and action, leading to the resolution of the issues that are being investigated (Stringer 2014). An action plan that resolves the problem at hand is therefore a commonly expected step in the methods based on action research methodology. However, this dissertation is a scholarly work in which a specific action plan is not expected or appropriate, and therefore the action plan step will be represented with case studies, which were used to test and improve the processes that

were developed as the research progressed. Three organizations served as case studies to demonstrate that SWI can be established through a specific process of engaging with the knowledge bases examined in this study.

- Montana State University Library
- McMaster University Library
- Coalition for Networked Information

These organizations were selected because a KC did not exist for them in initial Google searches, but they were also selected in part because they were convenient. *Montana State University Library* is where this research began, and the author and his colleagues collected evidence of its SWI condition when they became aware of the problem and along stages of development as they learned how to affect SWI. *McMaster University Library* was chosen because a colleague who is an Associate University Librarian (AUL) there volunteered his organization as a test case after hearing the author give a presentation on the topic in December 2014. The *Coalition for Networked Information (CNI)* represented a library professional organization that was also clearly struggling with its SWI, and the topic piqued the interest of the executive director when the author discussed it with him at a CNI membership meeting. The crucial convenience factor was that each of these three organizations provided the necessary account holder access that is required to claim or improve a business in GMB. Access of this kind is typically granted only in a trusted relationship, making it difficult to conduct these experiments on a large scale.

As further proof the effectiveness of the SWI process, evidence from before and after intervention was provided for three more organizations from Montana State University. This work was conducted by the author's colleague as part of a new SWI service being offered by the MSU Library, which will be discussed in more detail in Chapter 6.

Section 3.2.8 Take Action

For each organization, the following steps were taken:

1. Claimed and verified the business with Google My Business
2. Aligned Google+ profile with the organization, eliminating duplicates that existed
3. Published or improved a Wikipedia article for the organization
4. Verified generation of a DBpedia record if one did not already exist
5. Created or improved a Wikidata record for the organization

Section 3.2.9 Evaluate Action

Results were measured by again conducting searches and measuring improvements in the robustness of the KC. The timeline of these measurements varied, particularly in the early stages as the author and his colleagues were experimenting with different methods and knowledge bases. For instance, there was a period of several months following the publication of the Wikipedia article for McMaster University when the author was checking DBpedia almost every day for the appearance of a record, because the expectation at the time was that DBpedia had an influence on the generation of a KC, which has turned out to be false.

Section 3.3 Limitations of the Research Methods

Commercial search engine companies are secretive about the methods and algorithms they use to generate search results. The secrecy is due in part to intellectual property concerns, but also because giving away too much detail can lead to "black hat" techniques by website developers and SEO consultants, whose intent is to gain every advantage in attracting users to their sites. As a result, there is little published information from the search engines that indicate the sources from which they draw. Patents filed by these companies reveal general intent, but lack specificity. A patent filed by Microsoft Corp., for example, states that "the system may identify Wikipedia as having particularly strong and trustworthy entity information and may recognize various pages at that site as describing entities" (Li et al. 2013), but it makes no definite statements about Wikipedia or other sources from which Microsoft might draw to detect and establish entities for its Bing Satori knowledge graph. This limitation of proprietary information means that direct cause and effect cannot always be established.

Access to records in the proprietary knowledge bases examined in this study (GMB and Google+) is necessarily limited to authorized account holders. A successful search for an organization's name in GMB will lead either to a screen that indicates the business has already been claimed by someone else (see Figure 21 in Chapter 5) or to a screen that gives an authorized searcher the opportunity to claim the business. Claiming and verifying a business is a lengthy and multi-step process that includes responding to a postcard that Google mails to the physical address of the business. In some cases, a phone conversation

with Google can help, during which additional evidence of authorization may be requested. Basic profile information for the organization includes fields such as the Name, Address, Phone, Website, Organization Type, Hours and a free-text Introduction. Similarly, access to an existing Google+ profile is restricted to authorized account holders, and a basic Google+ profile is auto-generated from a claimed and verified GMB profile. Because of these restrictions it is impossible to evaluate the completeness of the record without authorized access, and this poses a limitation to the research method. Information missing from these profiles probably affects the presence and robustness of the KC, but this limitation could only be resolved in a study that includes account access to a large group of organizations.

Another limitation lies in the fact that the Semantic Web is a fluid environment where development and change are constant. The core knowledge bases that are addressed in this study seem relatively stable for now, notwithstanding the recent demise of Freebase, which had been an established source of significant information for Google's Knowledge Graph. In this case, at least, Google has decided to support the community-based efforts behind the Semantic Web by migrating its Freebase data into Wikidata (Tanon et al. 2016a). Recent reports indicate that Wikidata is playing an increasing role (Edward 2015) as a data source for Google's Knowledge Graph, and while it may not become as significant a source for the Knowledge Graph as Freebase was, it will probably play a role as one of several significant data sources.

Search engines are known to customize results based on cookies and the user's login and profile (Izenstark 2014), and therefore some effort was made in this study to minimize intrusions that could affect results based on user preferences, past history and location. The Google Chrome browser's "Incognito" was utilized for most searches (except for GMB and Google+), and this helped reduce some of these concerns, although it does not eliminate them.

Finally, while use of the Internet through mobile devices and applications continues to climb, and while SWI is at least of equal concern in the mobile environment, the author has chosen to focus on KC as they display in the desktop environment. The mobile environment is currently less established than the desktop environment, and there are two different major operating systems (iOS and Android) that would have to be considered. While a similar study for the mobile environment would be a natural progression of this research, it is considered too unwieldy for this dissertation and has been left for others.

Section 3.4 Summary of Research Methods

This chapter followed the action research methodology to describe methods that were employed to gather and analyze data for this study. Data for 219 primary and alternate names of the ARL libraries were gathered by conducting searches in Google and recording the presence or lack of a KC for each as well as the appearance of eight common information elements that could be displayed on the KC. Searches were also conducted in five knowledge bases (GMB, Google+, Wikipedia, DBpedia, Wikidata) to determine the presence or lack of records for the organizations. Evidence was collected by generating screen capture files for each search result. Data were analyzed using the R statistical computing software, and findings will be presented in the next chapter.

Chapter 4 Findings

Section 4.1 Introduction

This chapter describes findings from the main source of data collection and analysis for this dissertation: 125 member libraries of the Association of Research Libraries (ARL). Findings for each of the three research questions and the two sub questions are addressed. The chapter continues with the three case studies, describing action s that were taken to establish SWI for the two academic libraries and one library professional organization. The chapter concludes with a more cursory review of three more academic organizations at MSU that have benefited from the SWI service being offered by the MSU Library.

A brief review of the data collection and analysis methods is appropriate. The author conducted searches in the Google search engine in December 2015 to determine the presence or lack of KC for the 125 members of the ARL. Of the 125 libraries, 94 were found to also have alternate names and therefore 219 total library names were searched. Primary and alternate library name searches intended to discover records for the organizations in five knowledge bases (Google My Business, Google+, Wikipedia, DBpedia and Wikidata) were conducted from January – April, 2016. Screen captures were made for the results of every Google search, as well as for the results of searches in each knowledge base listed above. These screen captures resulted in more than 1400 files that provide evidence of findings. Screen displays were captured even made when the search produced null results, with one notable exception: DBpedia. When Wikipedia articles did not exist, a screen capture could be taken of the resulting message “The page X does not exist,” but since DBpedia records are generated from Wikipedia, the absence of a Wikipedia article meant no DBpedia URL existed that could be resolved into a page display and captured. As a result, many fewer DBpedia records are included in the data set than records for the other knowledge bases.

Data indicating the presence or lack of KC, presence or lack of records in the five knowledge bases, and scores for information elements, were recorded in an Excel spreadsheet. The spreadsheet was in turn converted into a comma delimited values (CSV) file that was read by RStudio, where error checking commands were run to validate the

data. Statistical analysis models and equations were then designed and run to answer the research questions.

The findings are presented, below, for each of the four research questions and for the two research sub-questions. For the first two research questions, a set of statements is presented, which are then supported or disputed by statistical analyses run through equations in R. Equations are listed and explained in Appendix B, along with numerical results. Graphical table plot figures are displayed in this chapter to help illustrate results of the equations and to supplement textual explanations. The figures that show table plots should be read as each row representing a primary or alternate name of a library, and thus there are 219 individual rows in each table plot. The presence of a value is indicated by a "1" while the absence of a value is indicated by a "0," and each table plot display has assigned colors to these values, which are explained in the captions below the figures.

Section 4.2 ARL Libraries Survey Findings

Section 4.2.1 Findings for RQ1:

What is the current state of Semantic Web Identity of ARL libraries, as indicated by the presence of accurate Knowledge Graph Cards in Google search results when the primary and alternate names of those libraries are searched?

All equations were run as pairwise relationship calculations of spreadsheet data columns to respond to the statements below. Each calculation is briefly discussed, and the equations themselves may be viewed in Appendix B along with results. Table 2 provides a numerical summary of findings for Research Question 1.

1) Number and percent of accurate KC found for either the primary or alternate names of ARL member libraries.
2) Number and percent of accurate KC for ARL libraries that displayed the same KC for both primary and alternate names.
3) Number and percent of primary library names that displayed accurate KC, and number and percent of alternate library names that displayed accurate KC.

Equation 1 shows the number of accurate KC that displayed for the 125 ARL libraries, regardless of whether the primary or alternate name was searched. The result shows that 102/125 (82%) ARL members displayed accurate KC, without distinguishing whether the KC displayed for the primary or alternate name of the library.

Equation 2 adds nuance to the first statement results by considering the issue of semantic "same as" relationships for primary and alternate library names. An established "same as" relationship would indicate that the search engine recognizes the primary and alternate names as belonging to the same library organization, causing it to display the same KC regardless of which name is searched. The result of the equation demonstrates that the "same as" relationship of primary and alternate library is lacking in most cases. Only 46/125 (37%) of ARL member libraries displayed the same KC whether the primary or alternate name was searched, putting into perspective the seemingly high outcome of the first equation. Figure 9 provides a visual display of this disparity.

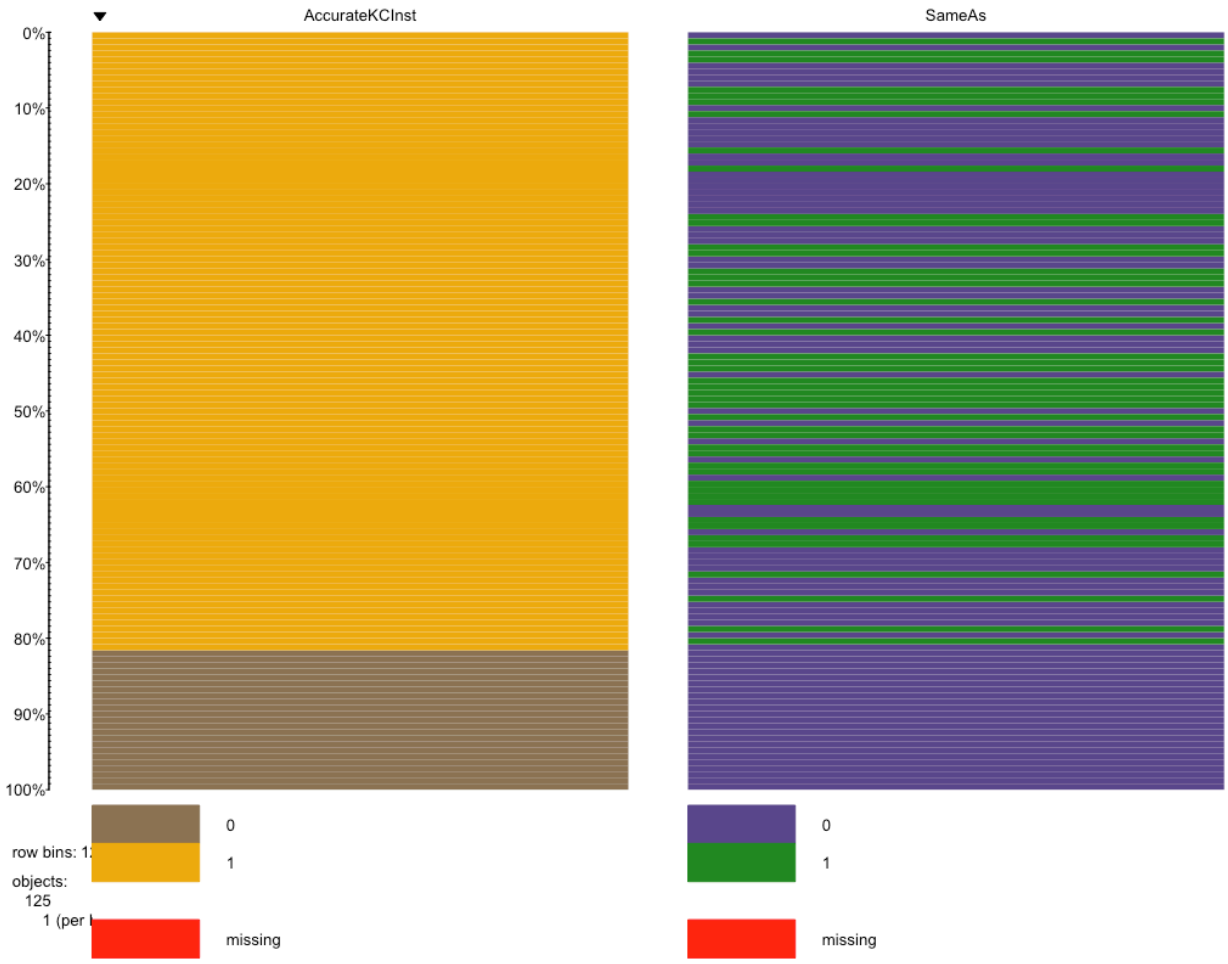

Figure 9: Table plot showing that 82% of ARL libraries displayed an accurate KC (Column 1, yellow rows), but that many of the KC were not the same for the primary and alternate names of the libraries were searched (Column 2, purple rows).

Equation 3 demonstrates the number of accurate KC that displayed for each library's primary name as well as each alternate name. Only 46% of primary names searched displayed accurate KC, while 79% of alternate names displayed accurate KC. Combined, 132/219 (60%) primary and alternate names displayed accurate KC, leaving 87 (40%) of primary and alternate names that either displayed no KC at all or displayed a KC that was inaccurate for the library name being searched. Figure 10 provides a graphical representation of this difference. As always, a value of "1" indicates presence, while "0" indicates lack of presence. In the left-hand column of the table plot "Primary=1" indicates a primary library name (blue color), while "Primary=0" indicates an alternate name (orange color). The right-hand column shows whether an accurate KC was displayed (green color) for each name. It is clear that alternate names (orange color) were more likely to display an accurate KC (green color).

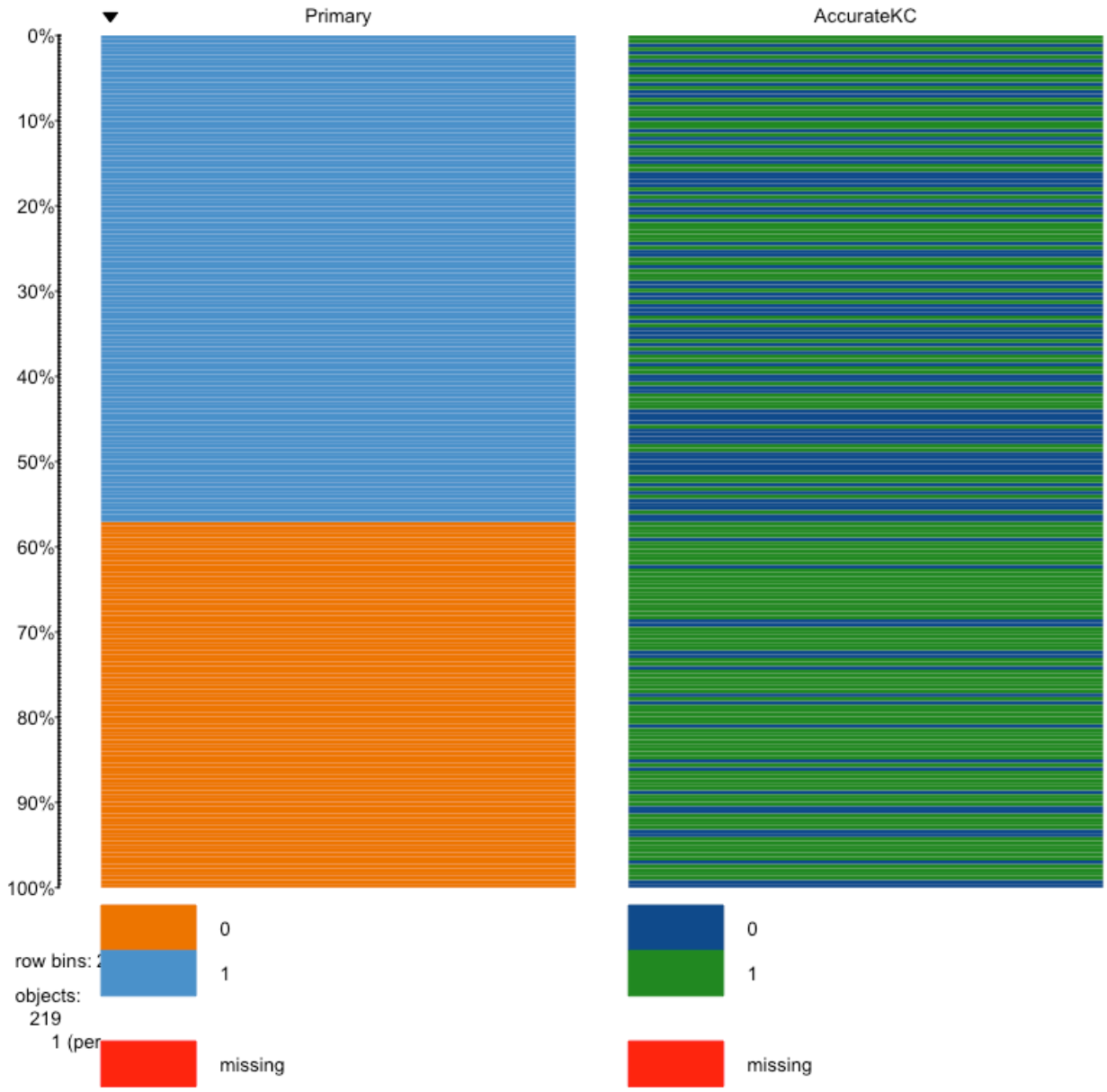

Figure 10: Table plot showing that ARL library alternate names (column 1, orange rows) were more likely to display an accurate KC (column 2, green rows)

Statements, R command strings, and responses for Research Question 1 are summarized in Table 2.

Statement	R command string	Response	Percent
Accurate KC displayed for either primary or alternate library name	t(with(SWI,table(PrimOrAltKC,AccurateKCInst)))	102/125	82%
Same KC is displayed for both primary and alternate library name	t(with(SWI,table(AccurateKCInst,SameAs)))	46/125	37%
Accurate KC displayed for total of primary and alternate names	t(with(SWI,table(Primary,AccurateKC)))	132/219	60%
Primary library names with accurate KC	(with(SWI,table(Primary,AccurateKC)))	58/125	46%
Alternate library names with accurate KC	(with(SWI,table(Primary, AccurateKC,)))	74/94	79%

Table 2: Responses to Research Question 1 (corresponds to Equations 1-3 in Appendix B)

Section 4.2.2 Findings for RQ2:

Are records or profiles present for ARL primary and alternate library names in the following knowledge bases: Google My Business, Google+, Wikipedia, DBpedia and Wikidata?

Arithmetic calculations were made to inform the following statements:

8) Number and percent of libraries that have "claimed and verified their businesses" in GMB
9) Number and percent of libraries that have unverified Google+ profiles
10) Number and percent of libraries that have verified Google+ profiles
11) Number and percent of libraries that have Wikipedia articles without infoboxes
12) Number and percent of libraries that have Wikipedia articles with infoboxes
13) Number and percent of libraries that have DBpedia records
14) Number and percent of libraries that have Wikidata records

Table 3 shows a summary of knowledge base records recorded for the primary and alternate names of each library organization. Equation 4 in Appendix B shows various

pairwise relationship equations were run in R to compare the spreadsheet column for each knowledge base with the spreadsheet column titled "Primary," which contained a "1" for the primary name of the library and a "0" for the alternate name. Since this research question was only designed to determine the presence of knowledge base records, reporting the presence of a KC was not relevant for this grouping. Percentages for the primary ARL library names were calculated from a possible 125, while percentages for alternate names were calculated from a possible 94 that were discovered during data collection. Total percentages were calculated from the sum of possible primary and alternate names, or 219. Percent figures were rounded up or down to the nearest whole number.

Knowledge Base	Primary (% of 125)	Alternate (% of 94)	Total (% of 219)
Google My Business	28 (22%)	40 (43%)	68 (31%)
Google Plus (unverified)	25 (20%)	17 (18%)	42 (19%)
Google Plus (verified)	22 (18%)	19 (20%)	41 (19%)
Wikipedia (w/o infobox)	10 (8%)	16 (17%)	26 (12%)
Wikipedia (w/infobox)	30 (24%)	26 (28%)	56 (26%)
DBpedia	30 (24%)	39 (41%)	69 (32%)
Wikidata	26 (21%)	37 (39%)	63 (29%)

Table 3: Number and percent of knowledge base records for primary and alternate names of ARL libraries

Section 4.2.2.1 Findings for RQ2, Sub-question 1

Is an accurate KC likely to display in search results if the library organization has not been claimed and verified in Google My Business?

Equation 5 in Appendix B shows the three-way relationship calculation that was run in R to reveal the following: 1) primary or alternate library names that; 2); had claimed and verified their businesses in GMB; and 3) displayed accurate KC at the time of data collection. Results show that 23/125 (18%) primary library names that displayed an accurate KC had a claimed and verified business profile in GMB, while 39/94 (41%) of alternate library names with an accurate KC also showed a claimed and verified profile in GMB. Conversely, 35/125 (28%) primary library names displayed an accurate KC without showing a claimed and verified profile in GMB, and 35/94 (37%) of alternate names also showed an accurate KC without a

claimed and verified business profile in GMB. This indicates that it is possible to have a KC without having claimed a business in GMB. However, it should also be noted that only 6 primary and alternate library names that showed a claimed and verified profile in GMB lacked accurate KC, while 81 primary and alternate names that lacked a claimed and verified profile in GMB also lacked an accurate KC. ARL libraries that have claimed and verified their businesses in GMB, therefore, are much more likely to display an accurate KC. This relationship is represented graphically in Figure 11.

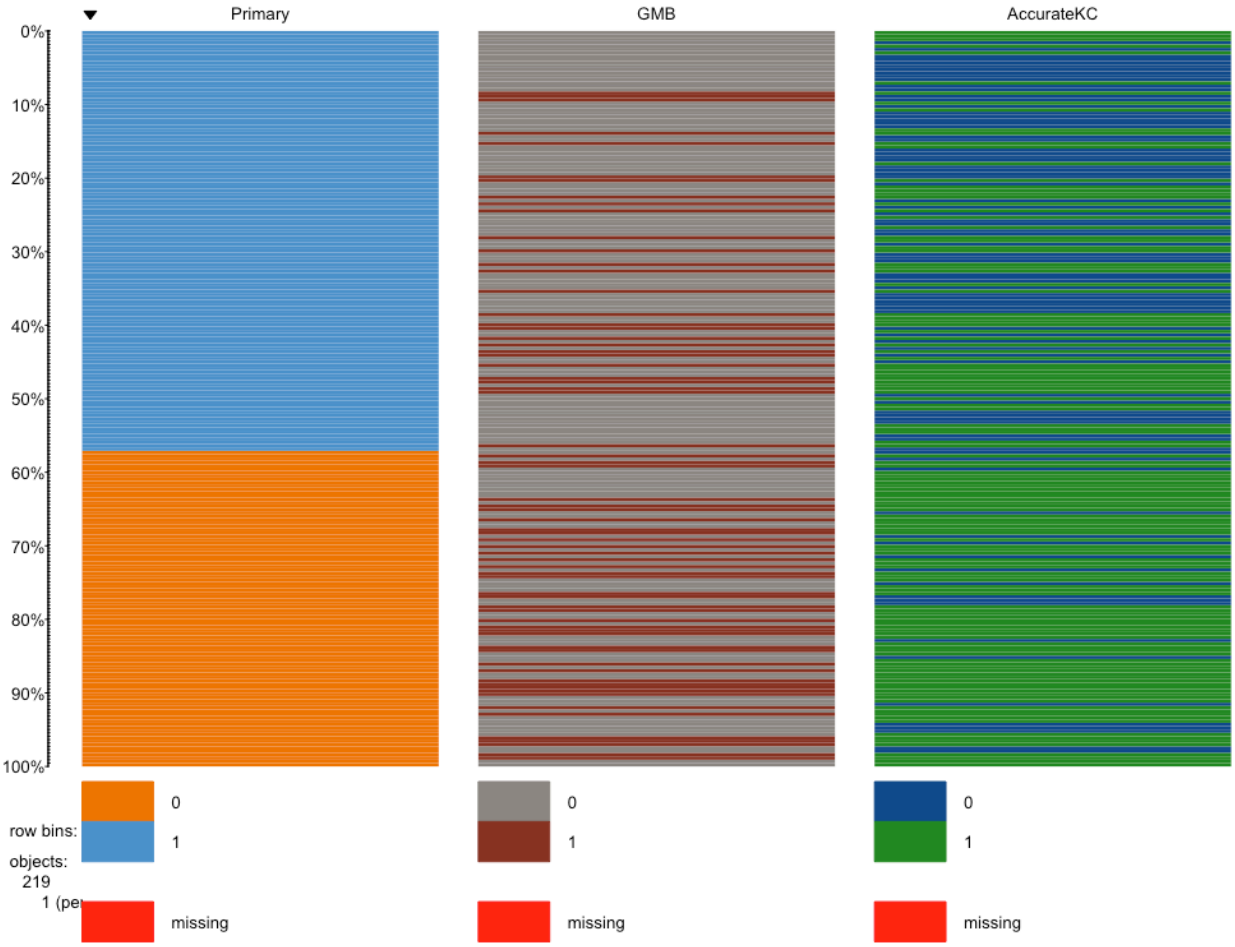

Figure 11: Table plot showing libraries that displayed a KC for their alternate names (column 1, orange rows) were more likely to have claimed a property in GMB (column 2, dark red rows) and were more likely to display accurate KC (column 3, green rows).

Section 4.2.2.2 Findings for RQ2, Sub-question 2

Is a KC likely to display a description field if a Wikipedia article does not exist for the primary or alternate name of the library?

To answer this question, a three-way relationship was calculated in Equation 6 that established: 1) the presence of a Wikipedia article; 2) the presence of a free-text description on the KC; and 3) the presence of an accurate KC.

- 60/132 library names (45%) that displayed an accurate KC had neither a Wikipedia article or a description visible on the KC.
- 47/132 library names (36%) that displayed an accurate KC also had a Wikipedia article and showed a description on the KC.
- Descriptions appeared on 10 accurate KC even though no Wikipedia article existed.
- 15 existing Wikipedia articles seemed to yield no description at all on the accurate KC.

Graphical evidence is shown in Figure 12 that articles published in Wikipedia tend to result in description fields on accurate KC

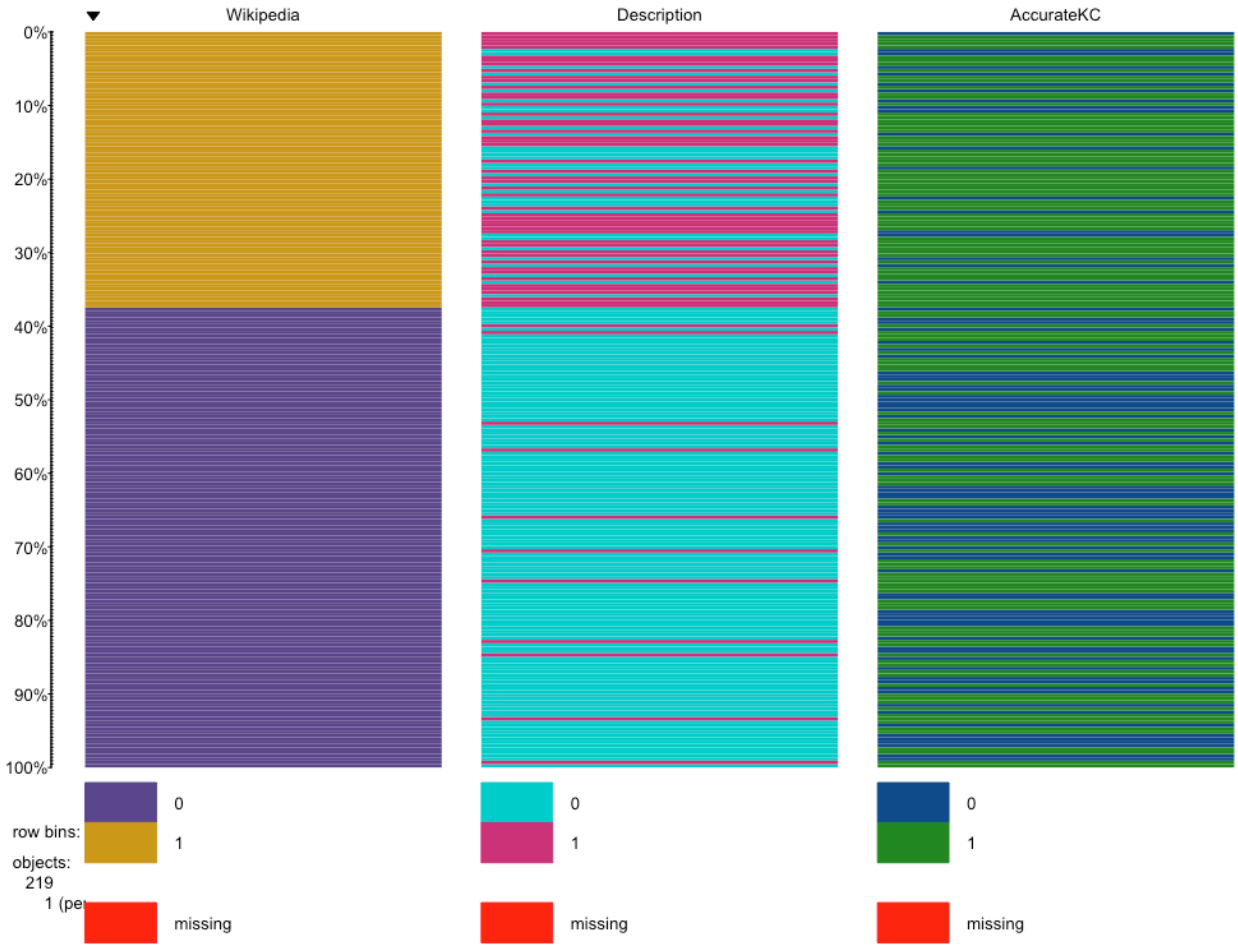

Figure 12: Table plot showing that Wikipedia articles (column 1, yellow rows) tend to result in descriptions (column 2, pink rows) on accurate KC (column 3, green rows).

Section 4.2.3 Findings for RQ3

Does the presence of a given knowledge base record predict the odds of information elements on the KC being populated?

Odds of information elements being affected by knowledge base records were estimated through logistic regression, which is a statistical analysis method that evaluates the relationship of the outcome variable (KC information elements, in this study) with the presence of one or more independent variables (knowledge bases, in this study). Logistic regression models are distinguished from linear regression models by the scoring of the outcome variable, which must be binary or dichotomous (Hosmer and Lemeshow 2000). For example, logistic regression is sometimes used to model predictions of medical conditions, where the binary outcome of diseased/non-diseased is modeled against various kinds of exposures (Dalgaard 2002). In this study, a simple binary score was recorded during the data gathering stage for the eight information elements, based on their presence (1) or lack (0) in the KC. This binary scoring necessitates the use of logistic regression over linear regression.

Although data were gathered for five independent variables (knowledge bases) only three (GMB, Wikipedia, and Wikidata) were used in the logistic regression equations. DBpedia and Google+ were eliminated as independent variables because of the statistical concept of multicollinearity, which warns against the possibility that one or more predictor (independent) variables are correlated by requiring that these "variables be truly independent of one another" (Farrar and Glauber 1967). Some level of multicollinearity is almost always present in similar studies, but substantiated conflicts in independence should be minimized, if possible. The literature shows that DBpedia records are only generated from existing Wikipedia articles (Morrison 2013; Lehmann et al. 2015; Bizer et al. 2009), and coupled with Google's pronouncement that its only relationship with DBpedia is through "transivity" (Mendes and Jakob 2012), the danger is high that the presence of a DBpedia record in this logistic regression could indicate correlation with Wikipedia rather than any independent predictive odds of DBpedia affecting the presence of KC information elements. Similarly, GMB auto-generates verified Google+ pages once an account holder has completed the GMB claiming and verification process (Carnduff 2014), and while it is possible for users to create Google+ pages independently of GMB, it cannot be stated with certainty that GMB and Google+ each bring different information into the model. For these

reasons of possible multicollinearity, DBpedia and Google+ were eliminated from the logistic regression models designed to respond to Research Question 3.

The outcome of logistic regression for this study predicts the expected change in odds of observing the presence of the grouped information elements, which is associated with exactly one change of level in an independent variable while holding the other independent variables constant. For example, Equation 7 is designed to predict the expected odds of the information element "Description" appearing in the KC if, for the library name in question, a business is claimed and verified in GMB, if an article is published in Wikipedia, or if a record is created in Wikidata.

The initial result of a logistic regression equation produces an estimate of log-odds coefficients. The coefficients are difficult to read and interpret, thus, they are converted to odd-ratios in R through another calculation called "exponentiation." The exponentiated coefficients show how much the odds are expected to increase "multiplicatively with a one-unit change in the independent variable" and "the dividing line between a positive and negative relationship is 1.0" (Drakos et al. 2005). In this study, the independent variables are categorical, meaning that changes move from one group to another (absent to present), rather than in one-unit linear changes within the independent variable as described by Drakos et al.

The "fit" command was run in R for each group of information elements, as demonstrated in Equation 7, where "Fit_d" establishes "Description" as a group (which, only for description, comprises a single information element). This equation is designed to estimate the odds of the three independent variables affecting an observed presence of a free-text description in the KC.

Section 4.2.3.1 Logistic Regression for the Description Group

Table 4 displays the log-odds coefficients of the predicted effect of the independent variables (GMB, Wikipedia, Wikidata) on the outcome group "description" in the KC. Log-odds coefficients will be shown only for this first logistic regression equation, since interpretations of prediction are generally drawn from the exponentiated coefficients.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5059     0.4011  -3.755 0.000174 ***
GMB1         -0.5821     0.4717  -1.234 0.217117
Wikipedia1    1.6914     0.7070   2.392 0.016735 *
Wikidata1     1.5735     0.7214   2.181 0.029166 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
```

Table 4: Log-odds coefficients of independent variables affecting description information element in accurate KC

For ease of interpretation, Table 5 shows the exponentiated coefficients as odds-ratios, combined into a single table that also shows confidence intervals. Confidence intervals are the range of plausible values for the change in odds of observing the presence of an information element group in the KC when moving from a library without a given independent variable (knowledge base record) to a library with that independent variable, while holding the other independent variables constant. The Intercept represents the odds of the outcome variable (in this case a description appearing on the KC) if none of the other three independent variables (GMB, Wikipedia, Wikidata) exist.

```
                OR          2.5 %       97.5 %
(Intercept) 0.2218283 0.09506305  0.4650701
GMB1        0.5587045 0.21835353  1.4119413
Wikipedia1  5.4271799 1.32303459 22.2804499
Wikidata1   4.8235735 1.17551736 20.8361664
```

Table 5: Exponentiated odds-ratios and confidence intervals of independent variables affecting the presence of the description outcome variable in accurate KC

The results in this case show the odds of a description field appearing in the KC to be well below 1.0 (0.222) for the Intercept, and the span of the confidence interval also well below 1.0. This indicates that the probability of observing the presence of a description in

the KC is smaller than the probability of not observing a description in the KC, assuming that a library does not have a GMB, does not have a Wikipedia page, and does not have a Wikidata page. The results also show the odds-ratio of .559 for GMB, with a 95% confidence interval for the estimate spanning 1.0 (.218-1.411). This indicates the model cannot estimate the change in odds of having a description in the KC as a result of a claimed and verified business in GMB with enough precision to statistically distinguish it from 1.0. In other words, the model can't tell whether GMB has a positive or negative effect on the odds of the presence of a description in the KC.

The presence of Wikipedia articles shows a positive odds-ratio (5.4) for the observation of the presence of a description. The confidence interval stays above 1.0, showing that it can be stated with 95% certainty that a Wikipedia article with or without an infobox increases the odds of a description element appearing in the KC by a factor of 5.4, assuming the other independent variables remain constant.

Wikidata also shows a strong relationship with the description element in the KC, with the odds of appearance of a description increasing by a factor of 4.8, with an associated 95% confidence interval of 1.18-20.83. In this case the model's best estimate for the multiplicative factor by which the odds increase is 4.8 when a Wikidata record exists, but taking into consideration the model's uncertainty, the actual factor lies somewhere between 1.18 and 20.83, with 95% confidence.

Section 4.2.3.2 Logistic Regression for Appearance Group

The Appearance group comprises the following information elements, whose presence may be observed in a KC: Image; Logo; Type. Equation 8 shows the R command string that was used to determine logistic regression for the Appearance group, while Table 6 shows the resulting odds-ratios with confidence intervals. Figure 13 provides a graphical representation of the observed presence of the Appearance group in association with claimed and unclaimed GMB properties.

```
                    OR     2.5 %     97.5 %
(Intercept) 5.3283299 2.3841522 13.655100
GMB1        2.2352887 0.7549311  7.544613
Wikipedia1  1.1982904 0.1830368 23.664316
Wikidata1   0.6751987 0.0337913  4.584527
```

Table 6: Exponentiated odds-ratios and confidence intervals for each of the independent variables affecting the presence of the Appearance group in accurate KC

The results show a high Intercept, indicating that the odds of observing the Appearance group in an accurate KC even without any of the independent variables has a factor of 5.3, with an associated 95% confidence interval of 2.4-13.7. GMB shows a lower odds-ratio of 2.23, with an associated 95% confidence interval of .75-7.5, indicating the model cannot show with any statistical certainty that GMB influences the Appearance grouping on accurate KC. The results are similar for Wikipedia, which appears to have a slight positive influence (factor of 1.2), although the confidence interval again betrays any certainty in even that slight prediction. The presence of a Wikidata record for the library name shows no positive prediction for the Appearance group on accurate KC.

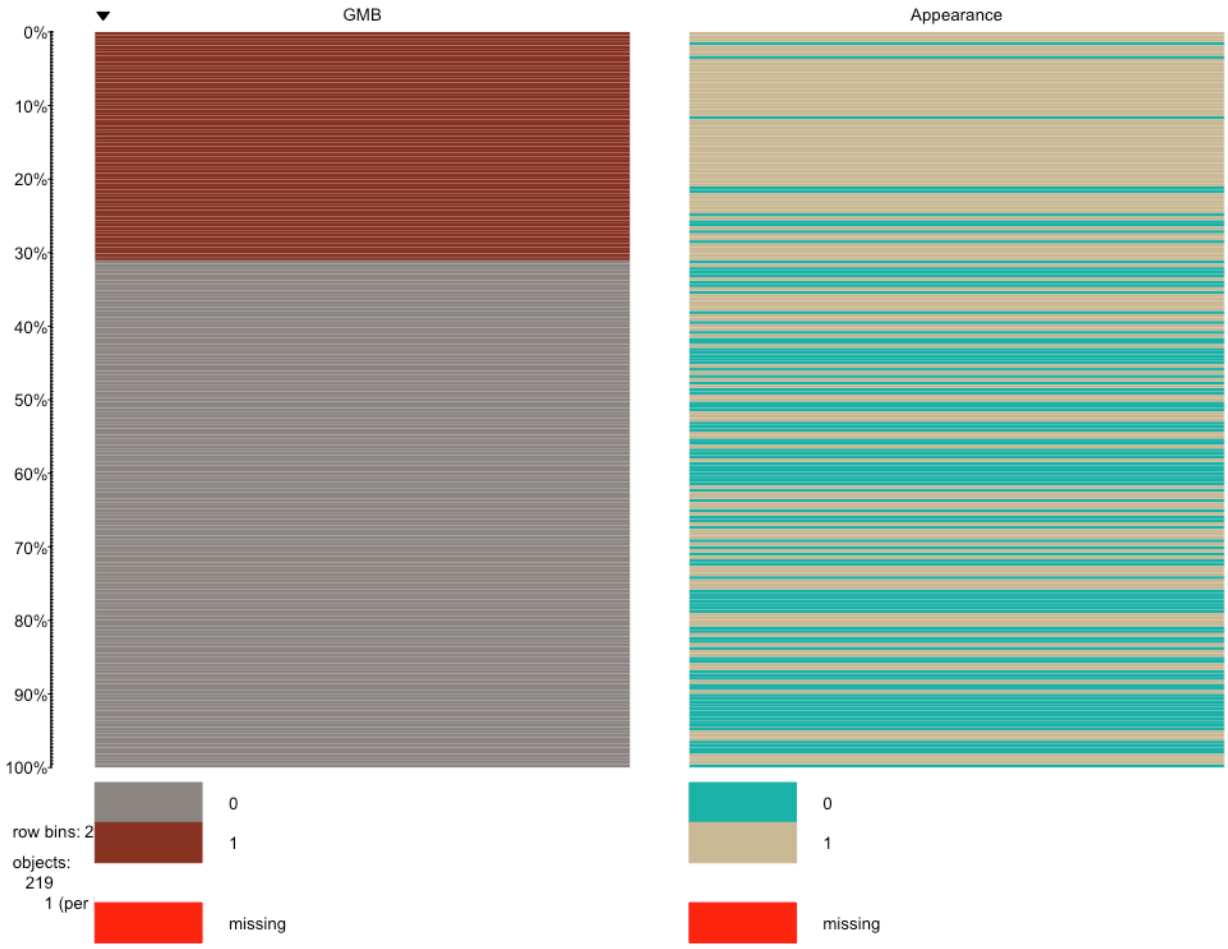

Figure 13: Table plot showing libraries that have claimed their property in GMB (column 1, dark red rows) are more likely to have KC with the Appearance group of elements (column 2, tan rows).

Section 4.2.3.3 Logistic Regression for Contact Group

The Contact group comprises the following information elements that may be present in an accurate KC: Address; Phone number; Website; Directions. Equation 9 once again shows the R command string used to determine logistic regression for the Contact group. Table 7 shows exponentiated coefficients as odds-ratios coupled with confidence intervals, and Figure 14 shows a graphical representation of GMB compared to the presence of the observed Contact group in KC.

```
                    OR      2.5 %    97.5 %
(Intercept) 9.2396773 3.68109561 29.054053
GMB1        2.0853251 0.69084950  7.159969
Wikipedia1  0.3175736 0.05621394  2.477503
Wikidata1   1.1337349 0.15210724  5.725517
```

Table 7: Exponentiated odds-ratios and confidence intervals for each of the independent variables affecting the presence of the Contact group in accurate KC

The Intercept again shows high odds of observing the elements that comprise the Contact group in the KC, with a factor of 9.2, supported by a strong confidence interval that indicates the factor could be as high as 29, without any of the other independent variables being present. Adding a GMB profile increases that factor by 2, but the confidence interval spanning 1.0 does not support this factor with any certainty. Wikipedia shows a weak predicted influence for the Contact group, while Wikidata shows a slight positive predicted influence, but again, the confidence interval spans 1.0, revealing that there isn't sufficient evidence to suggest that Wikidata affects the presence of the Contact group on the KC.

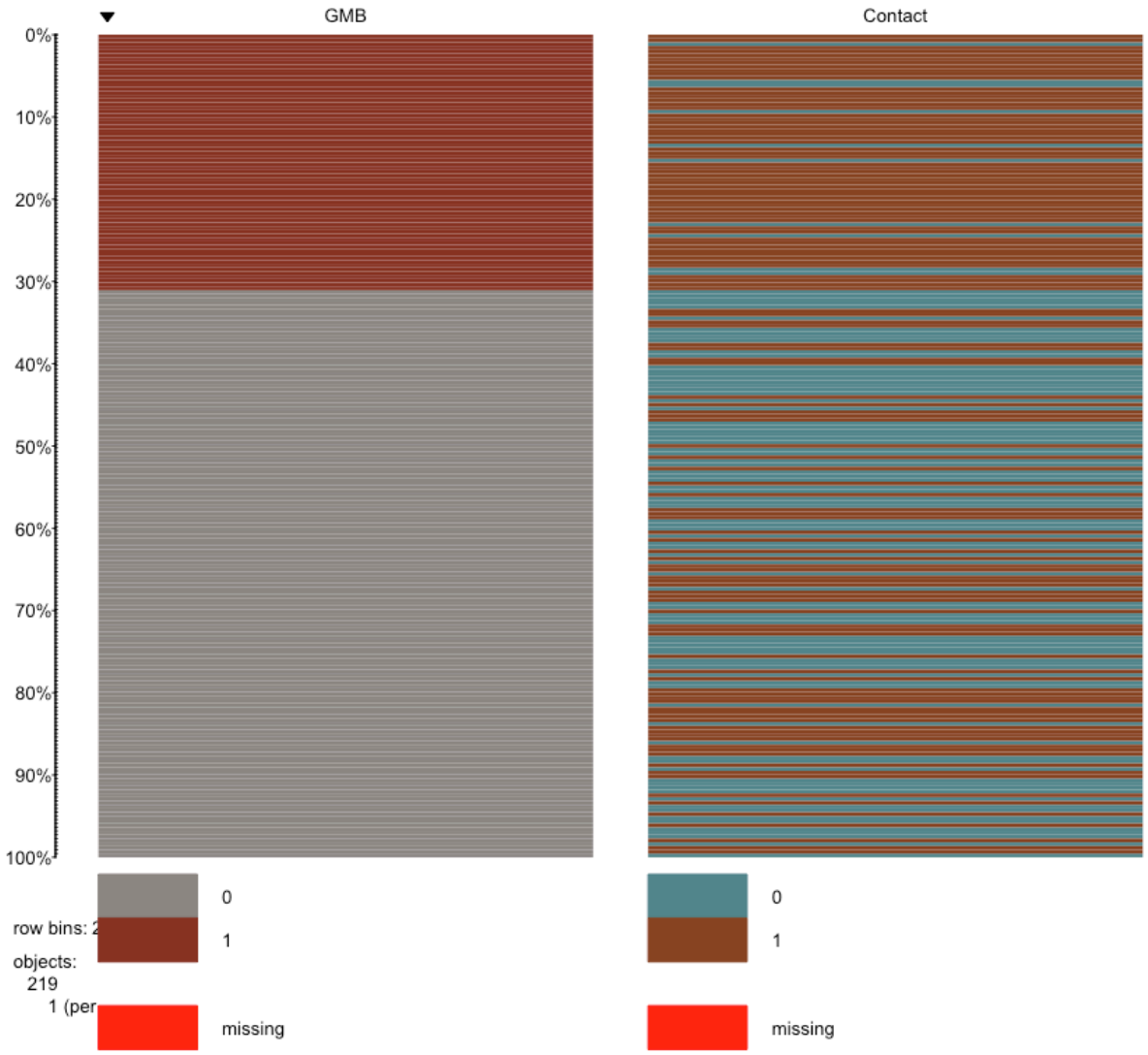

Figure 14: Table plot showing claimed GMB properties (Column 1, dark red rows) against observed presence of the Contact group in KC (Column 2, brown rows)

Section 4.3 Case Studies

Three organizations served as case studies to demonstrate the successful establishment of SWI during the period of 2013-16. The process for developing SWI was being studied and developed at the Montana State University Library throughout this period, and while the steps and timing varied slightly for each of the three organizations presented below, the overall process was similar for each. It should be noted that Freebase was an acknowledged source for Google's Knowledge Graph during the initial phase of the case studies, and was therefore part of the process with Montana State University Library. However, Google announced in 2014 that it planned to shut down Freebase and migrate its data to Wikidata (Google Knowledge Graph Team 2014), and therefore the subsequent case studies do not include any mention of Freebase.

Each of the three organizations is described, briefly, after which two tables are displayed for each. The first table summarizes the SWI conditions for the organization prior to any intervention, while the second table shows the actions taken and the results. Supporting screen capture examples may be found in Appendix C unless otherwise noted.

Section 4.3.1 Montana State University Library, 2013-16

MSU Library serves the students, faculty and surrounding community of Montana State University, a land-grant research university whose flagship campus is located in Bozeman, MT, USA. As mentioned previously, the author discovered in November 2012, that a Google search for "Montana State University Library" incorrectly displayed a KC for the library at a branch campus in Billings, MT instead of the flagship campus. Investigation by the author and his colleagues began shortly thereafter, and in early 2013 several MSU Library staff and faculty took the first formal steps that would eventually lead to dramatically improved SWI for the MSU Library.

Section 4.3.1.1 Summary of Conditions in January 2013

Knowledge Graph Card	Displayed the wrong organization and the wrong location
GMB	Business had not been claimed and verified
Google+	Two profiles existed, neither verified

Wikipedia	No article
DBpedia	No record
Wikidata	Wikidata was launched in October, 2012, and its use was initially very limited (Vrandečić and Krötzsch 2014). There is no reason to think a Wikidata record would have existed at the time, and a screen capture from September 26, 2013 confirms this (Figure 46).
Freebase	A record for *Renne Library* (the name of the library building on the Bozeman campus) had been created on March 10, 2012 by someone who was unknown to the MSU Library. No record was in evidence for *Montana State University Library.*

Section 4.3.1.2 Actions and results

GMB	Business was claimed and the record improved by August 2014.
Google+	Officially verified in August 2014, and second Google+ profile was successfully deleted.
Wikipedia	Research and writing began in the spring of 2013 under the guidance of experienced Wikipedia editors. The article was published on September 5, 2013.
DBpedia	Record appeared in April 2014 data release.
Wikidata	Record was generated from Wikipedia by a bot on November 29, 2013.
Freebase	A new Freebase record was auto generated on September 10, 2013, five days after the publication of the Wikipedia article. The record was titled *Montana State University Library* and was generated by a bot called "wikirecon_bot." MSU Library faculty added a "same as" declaration to the *Renne Library* Freebase record that linked it to the *Montana State University Library* Freebase record.
Knowledge Graph Card	An accurate KC began to appear for the MSU Library in September 2013, and gradually evolved to become much more robust as other knowledge bases were populated (Appendix D, Figure 72).

Section 4.3.2 McMaster University Library, 2015-16

McMaster University Library displayed no KC in December 2014 when the author conducted a first-pass survey of ARL libraries. Associate University Librarian, Dale Askey, approached the author following his presentation at the *Coalition for Networked Information* 2014 Fall Membership meeting (Arlitsch 2014b) and volunteered *McMaster University Library* as a case study to help develop the process that could establish SWI.

Section 4.3.2.1 Conditions in early 2015

Knowledge Graph Card	None (Figure 48)
GMB	Five different claimed and verified businesses existed for libraries on the McMaster University campus: 1. McMaster University Library – main library system 2. Mills Memorial Library – name of the building for the main library 3. H.G. Thode Library – Science and Engineering 4. Innis Library - Business
Google+	Verified and unverified profiles existed for at least three of the libraries
Wikipedia	No article (Figure 49)
DBpedia	No record
Wikidata	No record

Section 4.3.2.2 Actions and results

Google My Business	Deleted McMaster Libraries profile and tried to merge the McMaster University Library and Mills Memorial Library accounts. Thode and Innis profiles were managed by staff at those libraries and no attempt was made to change those.
Wikipedia	An article stub was first created in a Wikipedia "sandbox" on April 17, 2015, and was formally published on July 11, 2015 (Figure 52).
DBpedia	A record began to appear in Live DBpedia almost immediately after the Wikipedia article was published, but was only published in the DBpedia data dump around March 30, 2016.
Wikidata	A new record appeared on July 21, 2015 ten days after the Wikipedia article was published.
Knowledge Graph Card	A minimal KC began appearing in Google search results around June 9, 2015 (Figure 51), and evolved to eventually become much more robust in early February 2016 (Figure 53).

Section 4.3.3 CNI: Coalition for Networked Information, 2015-16

CNI is a "joint initiative of the Association of Research Libraries (ARL) and EDUCAUSE" that "promotes the use of digital information technology to advance scholarship and education." [4] As an institutional membership association, it offers strategic planning services, publications and reports, expertise, advocacy, and meetings that bring its members together for exchange of ideas and developments. CNI's SWI was a mixed condition when the author began to examine it in late 2015.

Section 4.3.3.1 Condition in late 2015

Knowledge Graph Card	None (Figure 54).
Google My Business	Two records existed; one complete and claimed business called "CNI" and the other was incomplete and was called

[4] https://www.cni.org/about-cni/history

	cnivideo.
Google+	Three profiles existed: one verified profile called *CNI* (likely generated from the claimed GMB profile of the same name); one unverified profile called *cnivideo* that had a YouTube channel connected to it; and a third, unverified profile also called *cnivideo* that contained almost no information. A search in Google+ for "Coalition for Networked Information" showed no Google+ profile at all (Figure 56)
Wikipedia	An article had been created on December 5, 2008, but in 2016 it still lacked an infobox and was flagged for not containing appropriate references (Figure 59).
DBpedia	A record with minimal information existed.
Wikidata	A bot-generated record had been created on February 22, 2013.

In December 2015 the author received permission from CNI's executive director, Clifford Lynch, to begin working with CNI staff to improve the condition of the organization's SWI. The author proposed a plan and was granted access to the organization's Google and YouTube accounts. He communicated with the Communications Coordinator at CNI, Diane Goldenberg-Hart, in advance of each step and documented changes.

Section 4.3.3.2 Actions and results

Google My Business	• Transferred ownership of GMB *cnivideo* profile to *CNI:Coalition for Networked Information* profile. • Deleted *cnivideo* profiles from GMB and Google+ while logged into the *CNI:Coalition for Networked Information* profile in GMB. • Prepended "CNI:" to the business name in the verified GMB profile on March 10, 2016, so that it became *CNI: Coalition for Networked Information* to match the

	organization's website.
Wikipedia	Added infobox to Wikipedia article on December 22, 2015 (Figure 60). There was no effect on KC, and the action came too late to be included in the DBpedia data dump from October 2015 that was released in April 2016.
YouTube	Transferred ownership of YouTube "cnivideo" channel from secondary account to the main CNI account on March 10, 2016, a two-step process that required acknowledgement from CNI staff. Completed transfer of channel on March 12.
Knowledge Graph Card	An accurate KC began appearing in Google search results on March 12, 2016 (Figure 61), but it appears consistently only when the full name of the organization is searched, i.e. "CNI: Coalition for Networked Information." At this writing (October 2016) the description field is still not appearing on the KC, which may be due to the fact that the Wikipedia article is still flagged as needing additional citations.

Section 4.3.4 Results from Additional Organizations

Several other organizations at Montana State University have received SWI treatment and shown similar results as the three case studies listed above. Work with the three organizations listed below represents the first phase of a planned SWI service for the entire campus, which is described in more detail in *Chapter 6*. Although the work with these additional organizations was performed by the author's colleague, and not the author himself, it nevertheless provide additional confirmation that the process is successful in establishing SWI. The additional organizations at MSU are:

1. Jake Jabs College of Business and Entrepreneurship (JJCBE)
2. Honors College
3. Office of the Provost

Each of these organizations at MSU lacked a KC prior to treatment, none had claimed their business in GMB, and none had Wikipedia articles, DBpedia records, or Wikidata

records. Work with the organizations began in the fall of 2015, and by mid-2016 each was displaying a KC. As before, steps taken to achieve this included claiming and verifying the business in GMB, creating a Wikipedia article, and creating or populating a Wikidata record. In some cases, such as with the JJCBE, an existing profile in GMB created by an anonymous user had to be merged with a new profile, which required some negotiation through the GMB Help Center[5].

Section 4.4 Summary of Findings

This chapter presented a statistical analysis of the data collected for ARL libraries, which demonstrated the condition of Semantic Web Identity (SWI) for the 125 members. The findings also showed results of the logistic regression efforts to make predictions about which knowledge bases influence the presence and population of Knowledge Graph Cards (KC) and some the information elements that populate them. It also showed results of intervention in three case studies by describing the existing SWI condition and the steps taken to improve it at *MSU Library*, *McMaster University Library*, and the *Coalition for Networked Information*. Three additional organizations at MSU were also presented, with brief descriptions of the SWI process that has been applied to them over the past year by the author's colleague, which essentially confirms the effectiveness of the process developed with the case studies. The next chapter will offer a more detailed discussion of the meaning and importance of the findings.

[5] Google My Business Help Center - https://support.google.com/business/?hl=en#topic=4539639

Chapter 5 Discussion

Section 5.1 Introduction

This chapter offers further analysis of the findings from the previous chapter. Findings for each of the research questions are discussed in depth and include screen capture figures to help illustrate results. A general discussion of each of the knowledge bases is included, as well as a discussion of other factors that influence the establishment of SWI. These include the effect of primary and alternate name use among ARL libraries and the problem of physical addresses in academic institutions.

Section 5.2 Analysis of Findings for the Research Questions

Section 5.2.1 Research Question 1

Data collected for this study demonstrate that there is room for significant improvement in the current state of SWI among ARL libraries, as measured by the presence or lack of accurate KC. The results also bring into focus the semantic difference between the concept of a thing (entity) and the name by which that thing must located in a search engine. It can be stated that 82% of the ARL library organizations (things) displayed an accurate KC, but the display of a KC is dependent on the name of the library that is searched. Google searches for the primary names of library organizations yielded accurate KC just 46% of the time, while the search for alternate names showed accurate KC 79% of the time. Combined, only 60% of the possible 219 primary and alternate names of the libraries displayed an accurate KC during Google searches. There is a discrepancy in the official names ARL libraries provide for lists like the ARL membership directory and how they represent themselves on the Web.

The analysis becomes much more interesting with the introduction of spreadsheet data that recorded the “same as” relationship when two KC displayed for a given ARL member library. If the search engine understands that an organization has two different names, the same KC would be displayed regardless of whether the primary or the alternate name was being searched. However, the data show that only 37% of ARL libraries enjoy this “same as” status. When accurate KC display for both the primary and alternate name of a

member library the two KC are usually not the same. In addition, the two different KC often display different facts about the same organization.

It is possible that Google may eventually determine the relationship of the primary and alternate names organically, by continuing to gather more information from websites and gradually building up the facts it has about an organization in its Knowledge Graph. However, a more reliable strategy is for an organization to explicitly establish the "same as" relationship in appropriate knowledge bases. GMB and Wikidata both have facilities for establishing the "same as" relationship (as did Freebase, when it was still a viable source). Wikipedia can also facilitate alternate names in its infobox templates.

Approximately 11% of KC that displayed during searches were for the wrong organization, and were therefore classified as inaccurate. In most cases, this inaccuracy was recorded because a KC displayed for a branch library on campus rather than the main library or the library umbrella organization that was being searched. For example, one would expect the name "Boston College Libraries," to be the umbrella name for all the libraries at Boston College, but using that search term in Google resulted in the display of a KC for the "Babst Art Library" (see Figure 15). Similarly, a search for "Yale University Library" displayed a KC for Yale's "Divinity School Library" (see Figure 16).

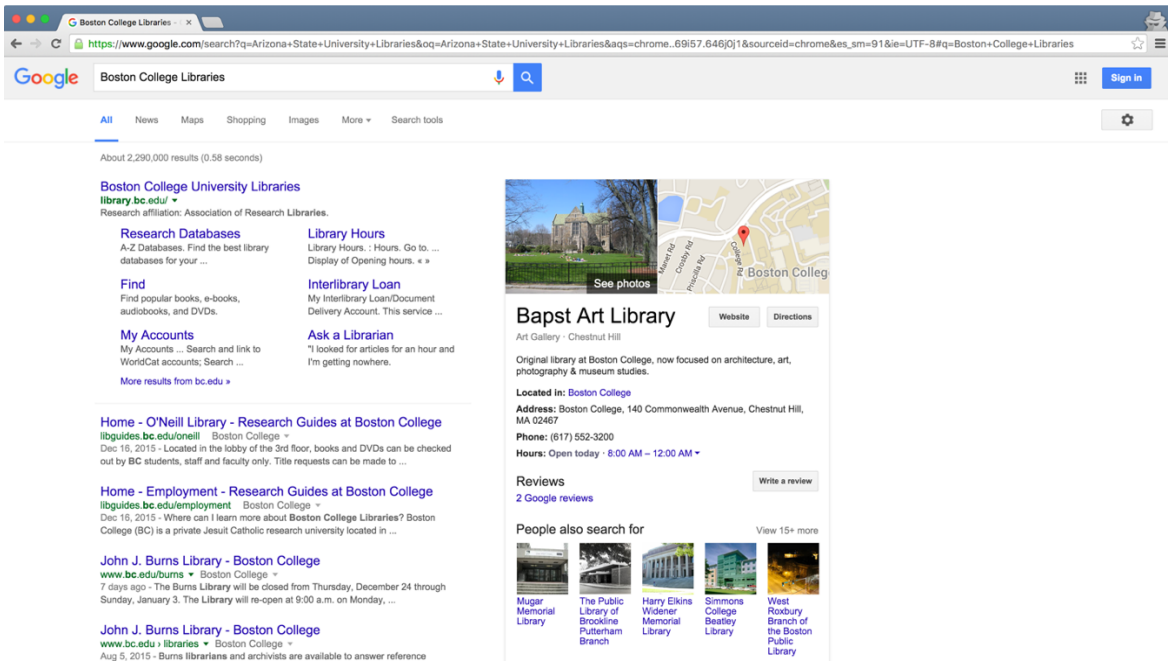

Figure 15: Google search for Boston College Libraries displayed KC for Babst Art Library at Boston College

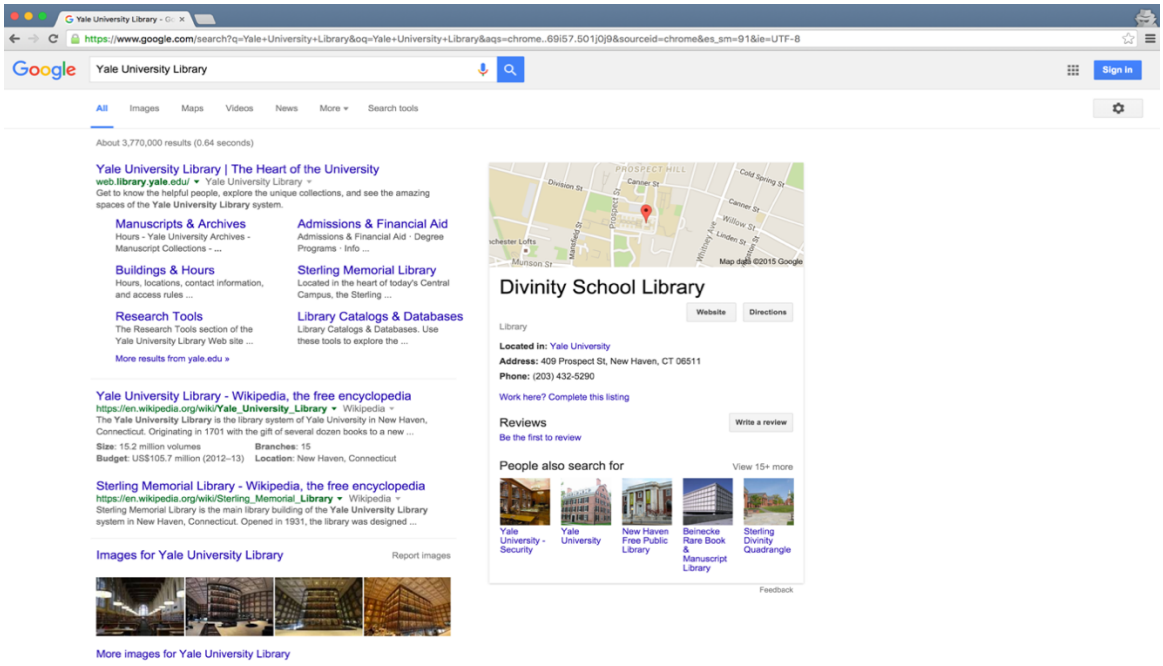

Figure 16: Google search for Yale University Library displayed a KC for Yale's Divinity School Library

In some cases, an inaccuracy was recorded because the KC showed an organization that was not affiliated with the campus libraries at all. For example, the search for the primary name "University of North Carolina Chapel Hill Libraries" displayed a KC for the "School of Library and Information Science" at UNC-Chapel Hill (see Figure 17).

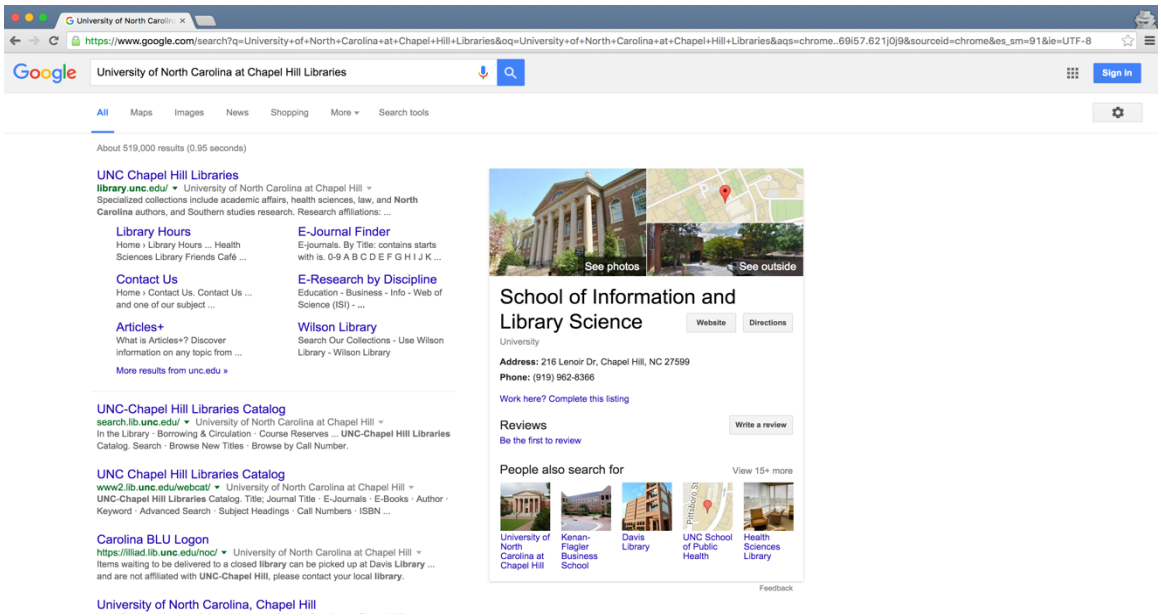

Figure 17: Google search for University of North Carolina at Chapel Hill Libraries displayed a KC for that university's School of Library and Information Science

Section 5.2.2 Research Question 2

Research Question 2 sought to determine whether records for the primary and alternate names of the ARL libraries could be found in the five knowledge bases that were tested. The finding of poor overall SWI of ARL libraries prompts the expectation that knowledge base records for the libraries would also be lacking, and in general this is supported by the data. The results of the searches for library primary and alternate names in the knowledge bases were described in Table 3 in Chapter 4. A greater number of records were found in GMB, DBpedia, and Wikidata for the alternate names of the libraries, while the records were evenly split between primary and alternate in Wikipedia, and tilted slightly in favor of primary names in Google Plus.

Section 5.2.3 Discussion of Knowledge Bases

The following section discusses each knowledge base that was examined for this study, including further explanation for some of the processes that result in records or profiles in each.

Section 5.2.3.1 Google My Business (GMB)

Findings from this study indicate that GMB may be the data source that has the single largest influence in helping establish SWI for organizations, as indicated by the number of accurate KC that are associated with established records in GMB. Establishing a record for the organization in GMB involves a claiming and verification process, including a communication exchange that helps Google assure the veracity of the claim. This process is crucial for establishing accuracy and trust in the information about the organization that is gathered into Google's Knowledge Graph.

While anyone can search GMB to determine whether a profile has been claimed and verified (see Figure 21), claiming and verifying a business, or viewing the details of the profile, requires a Google Account. Ideally that account is created on behalf of the organization and is shared by several administrators rather than being owned by one individual who may eventually leave the organization. Following the initial creation of a profile in the GMB system, Google sends a postcard to the physical address of the organization as entered by the claimant, which may seem archaic considering the advanced

information ecosystem the knowledge base is intended to support, but it is apparently the most effective method of crossing the lines from the physical world to the digital. The claimant must respond to the postcard within a given period, after which GMB will mark the business as "claimed" in its public interface and will generate a verified Google+ profile with the information that was entered into the GMB profile. In some cases, GMB may allow the verification process to take place by telephone, but then they require some additional evidence, such as external photographs of the building that the caller claims to inhabit.

A completed profile includes many of the information elements that can be found on a KC, including the organization name, address, phone number, website, and photographs. Once a business is claimed and verified, GMB can provide certain data "insights" to account holders that monitors search behavior and traffic associated with the organization. These include information about whether customers searched directly for the organization by name or whether they searched by category, product, or service. Insights also include the actions that users took once they viewed the KC, including visiting the website, requesting directions, calling by telephone, or viewing photos. These data points are collected by Google when a user clicks the information elements on the KC.

Claimed and verified businesses were evident for only 22% of ARL libraries' primary names, while 43% of ARL alternate names had been claimed. The higher number for the alternate names supports the proposition that ARL libraries suffer from a lack of consistent reference to their organizations, and that libraries use their primary names less on the Web than their alternate names. Combined, only 66 of 219 (30%) ARL libraries had claimed and verified their business in GMB for either their primary or alternate names. Figure 18 provides a visual display of this result, and shows that a library that has claimed and verified its property in GMB is more likely to display an accurate KC.

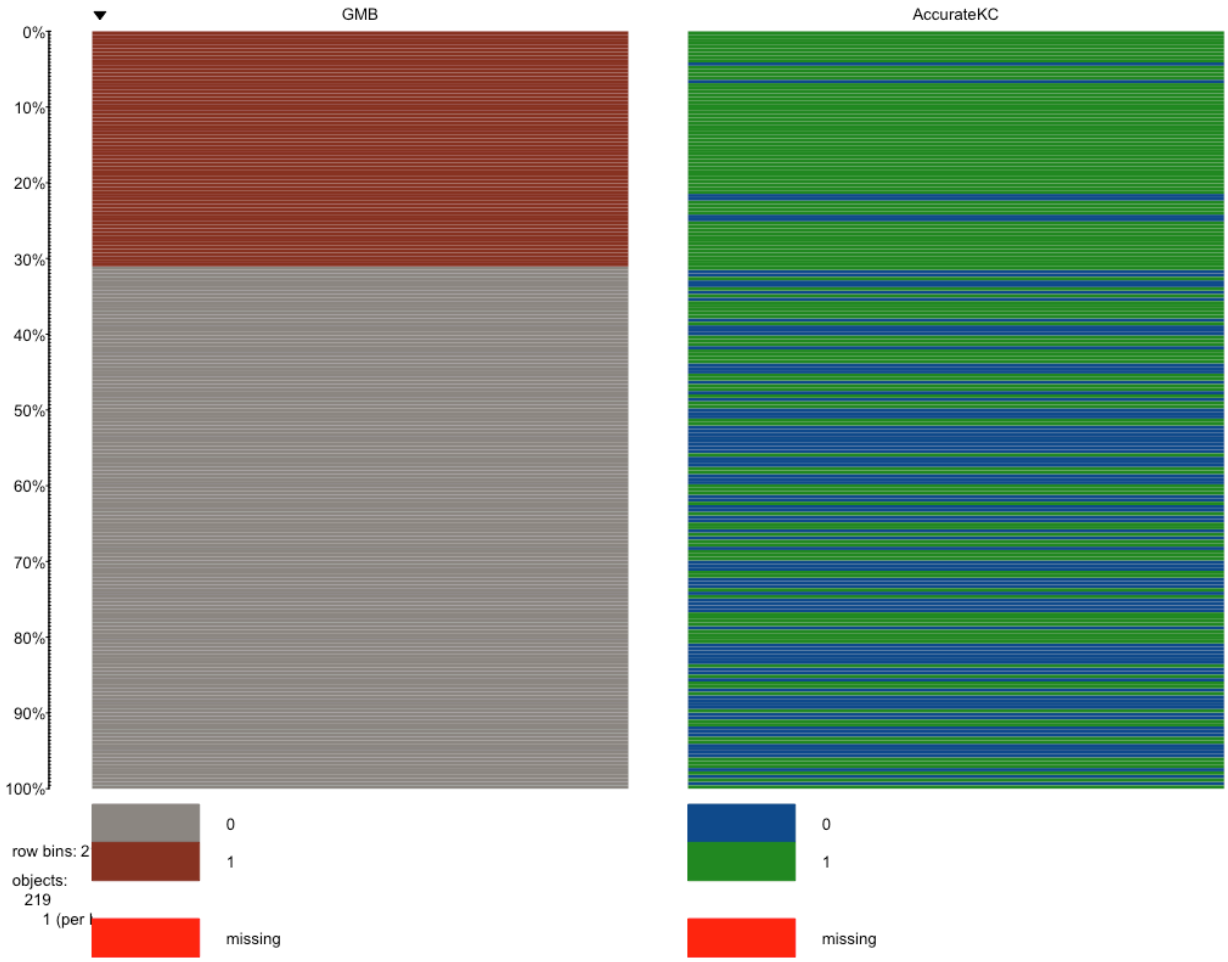

Figure 18: Table plot showing that the libraries that have claimed and verified their businesses in GMB (left column, dark red rows) are more likely to display an accurate KC (right column, green rows)

Section 5.2.3.2 Google Plus

The launch of Google+ in 2011 was met with anticipation by many Internet users, not only because of the obvious challenge to Facebook's dominance in the social media environment, but also because of Google+'s recognized advantage "in that it shares, communicates with and enhances already popular Google services" (Elson Anderson and Still 2011). Google+ is actually only the latest of Google Inc.'s forays into social media networking; earlier attempts that the company launched and eventually abandoned included Orkut in 2004 and Buzz in 2010 (Ovadia 2011; Jackson 2010; Orsini 2014). The timing of the Google+ launch meant it also faced immediate competition from other developing social platforms, such as Twitter, Tumblr, Instagram, SnapChat, etc. It soon became apparent that it would not succeed in displacing Facebook as the social networking site of choice, but Google+ has continued to evolve and its value may yet be realized. Its

integration with an array of Google properties gives it potential that other social media sites cannot match, and chief among these is the ability of Google+ profiles to surface prominently in Google SERP.

As with GMB, ARL libraries engagement with Google+ is limited and sporadic. The results from this study show that only 83 of 219 ARL primary and alternate library names (38%) had a Google+ profile in 2016. Of these 83 profiles, only 41 (19%) were verified by Google. The bar chart in Figure 19 shows the number of Google+ profiles for primary and alternate ARL library names, as well as the number of verified and unverified profiles. A discussion of the reasons for verified and unverified profiles follows in the next paragraph.

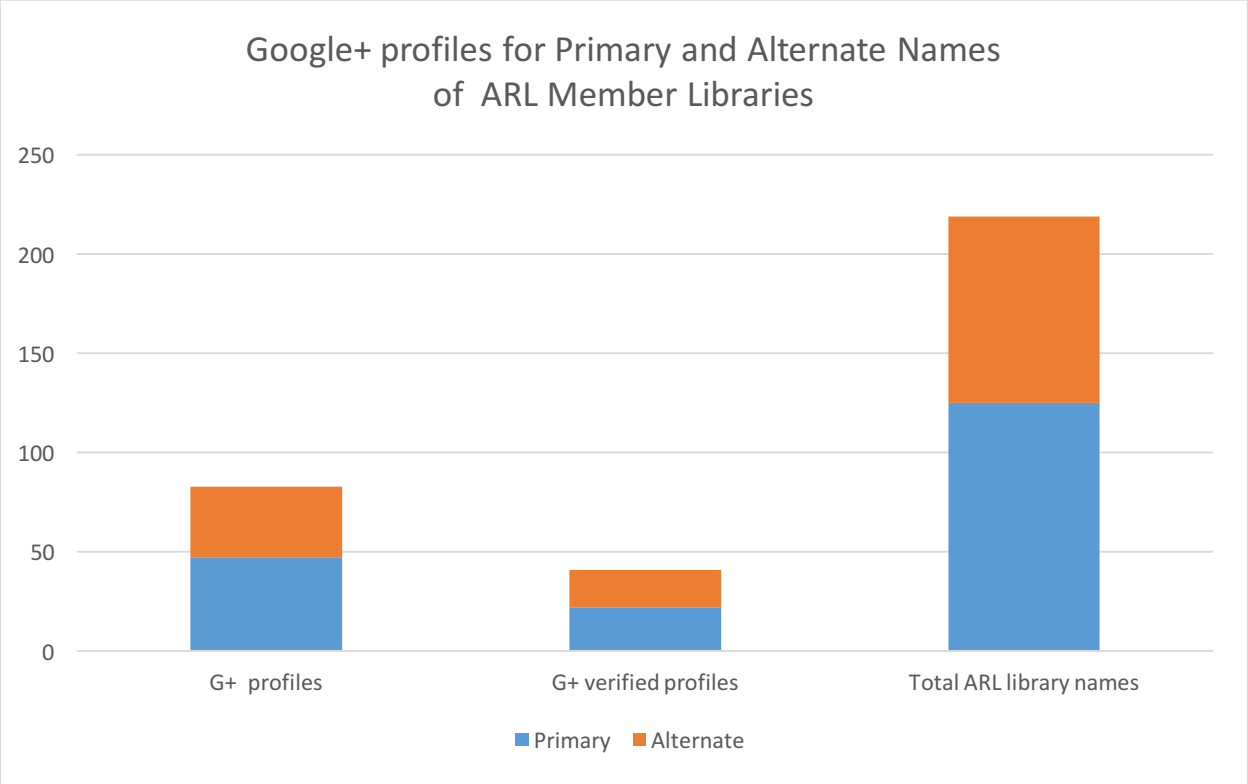

Figure 19: Chart showing libraries that have verified and unverified Google+ profiles for their primary and alternate names

Successfully claiming and verifying a business in GMB automatically generates a verified Google+ profile that is populated with basic information about the organization. The account holder can then supplement this information with additional photographs, posts to communicate with users, and a connection to the organization's YouTube channel. However, it is also possible for individuals to create Google+ profiles independently of the GMB business-claiming process; in fact, half of the Google+ profiles found for ARL libraries

were unverified. These were most likely created by individual employees without direction or support from administration, and some speculation about the reasons for this may be warranted. In the absence of administrative directives for marketing the organization on the Semantic Web it's easy to imagine employees who are more familiar with social media platforms having become impatient and taken matters into their own hands.

The University of Houston M.D. Anderson Library, for instance, has three Google+ profiles: one verified and two unverified. The verified profile (Figure 20) was apparently generated from the claimed GMB profile of the same name (Figure 21) and shows a robust profile, but the two unverified profiles are sparsely populated (Figure 22 and Figure 23). A fourth Google+ profile, for "University of Houston Libraries" also exists and is unverified (Figure 24). None of the four profiles displays the same physical address, which may contribute to the fact that two different KC appear in SERP with different addresses and different telephone numbers (see Figure 25 and Figure 26).

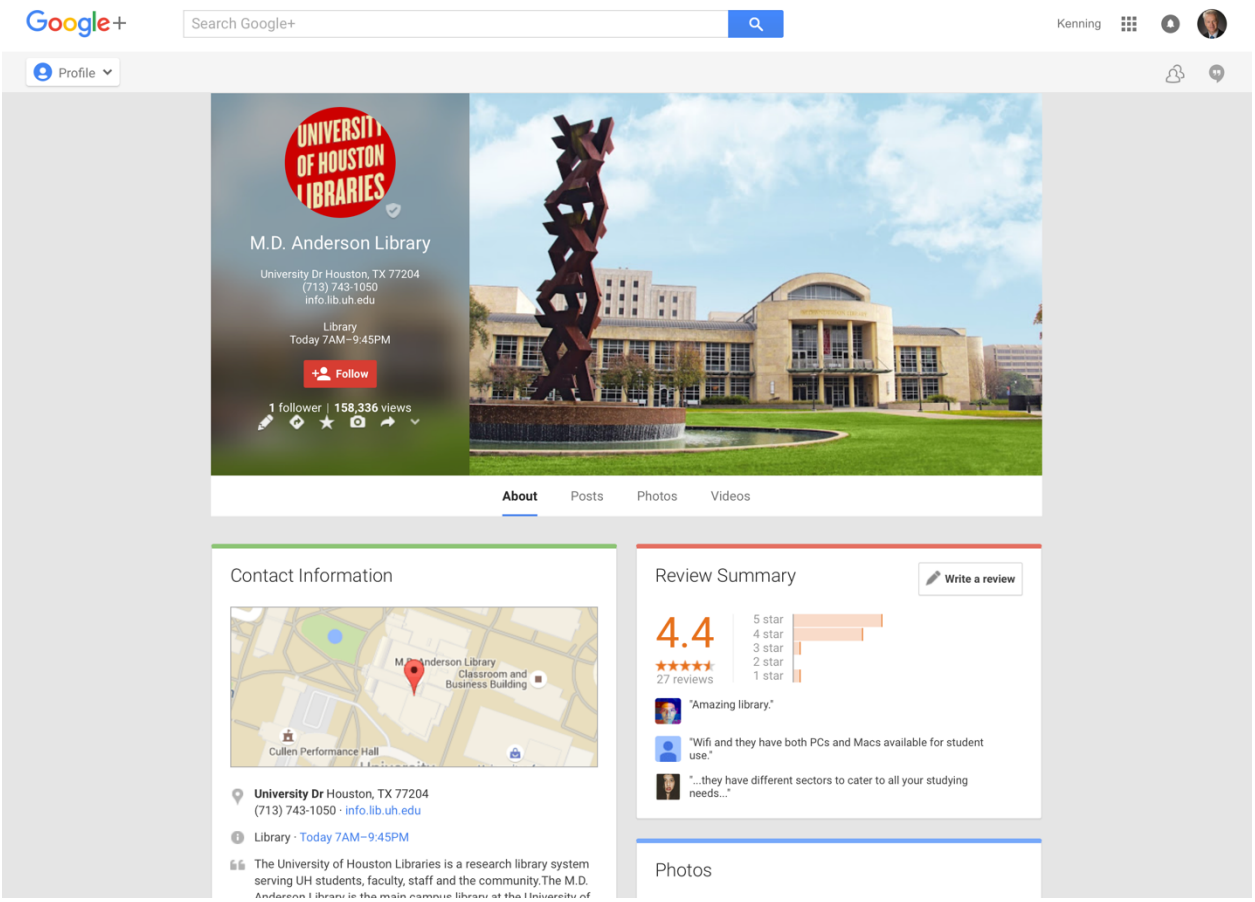

Figure 20: Verified Google+ profile for M.D. Anderson Library

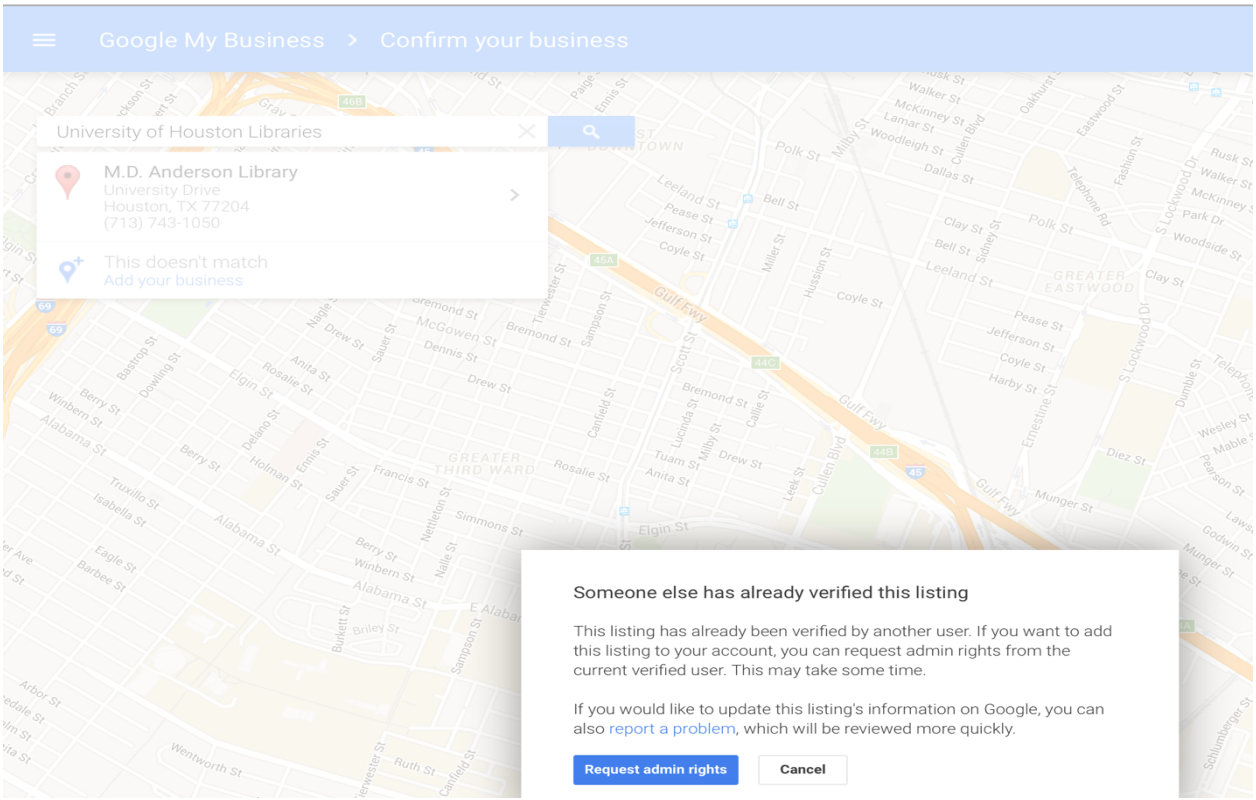

Figure 21: Claimed business name and address in GMB match Google+ profile in Figure 5

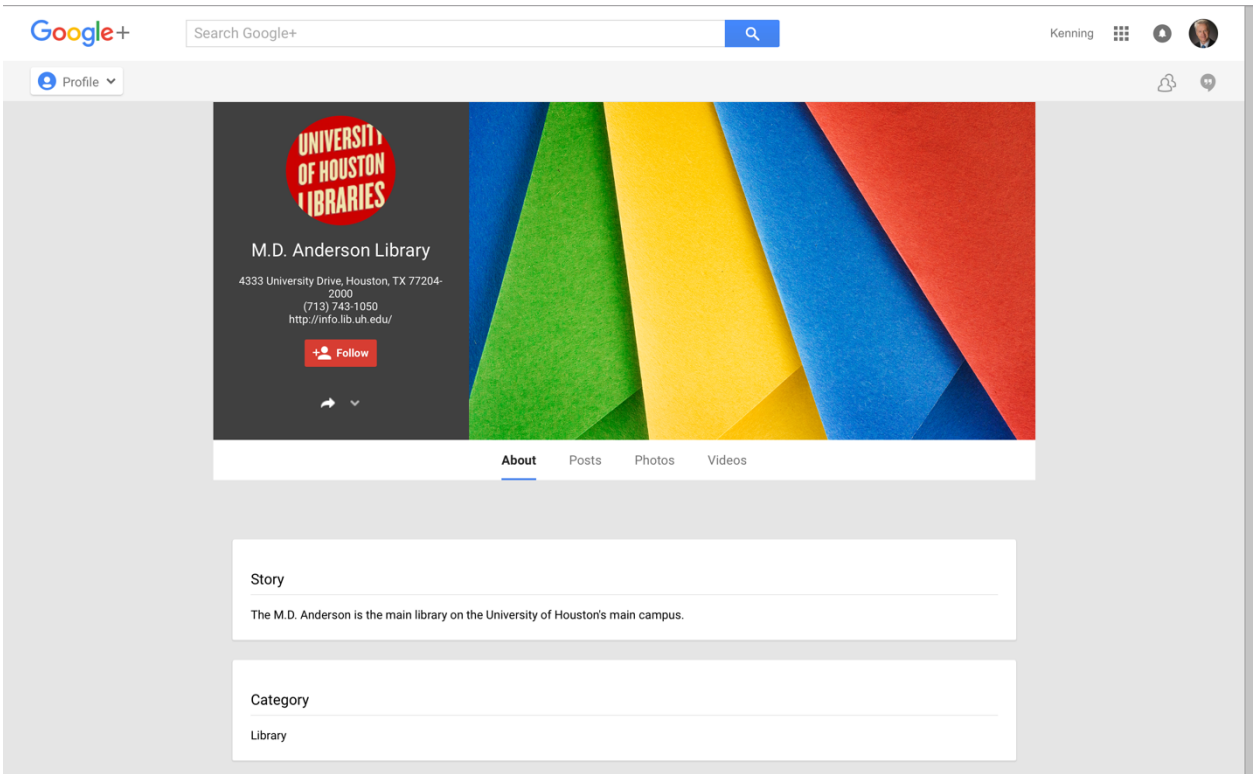

Figure 22: Unverified Google+ profile M.D. Anderson Library

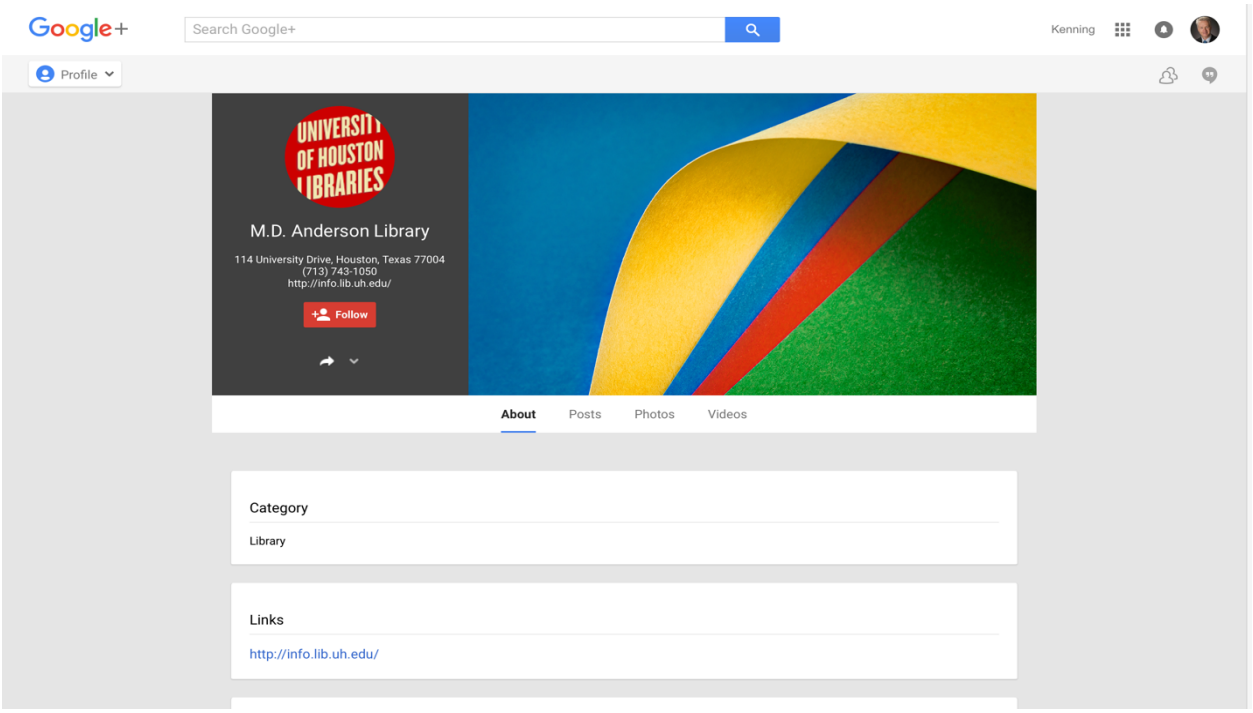

Figure 23: Second unverified Google+ profile for M.D. Anderson Library

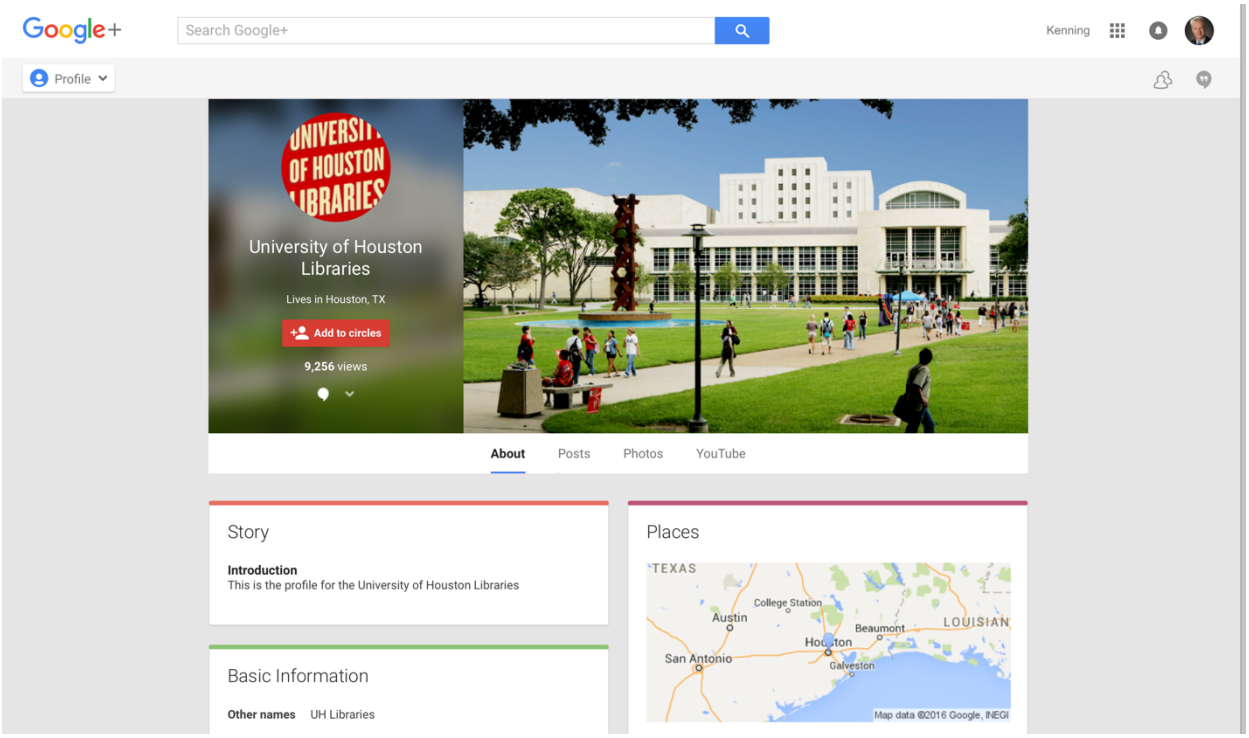

Figure 24: Unverified Google+ profile for the University of Houston Libraries

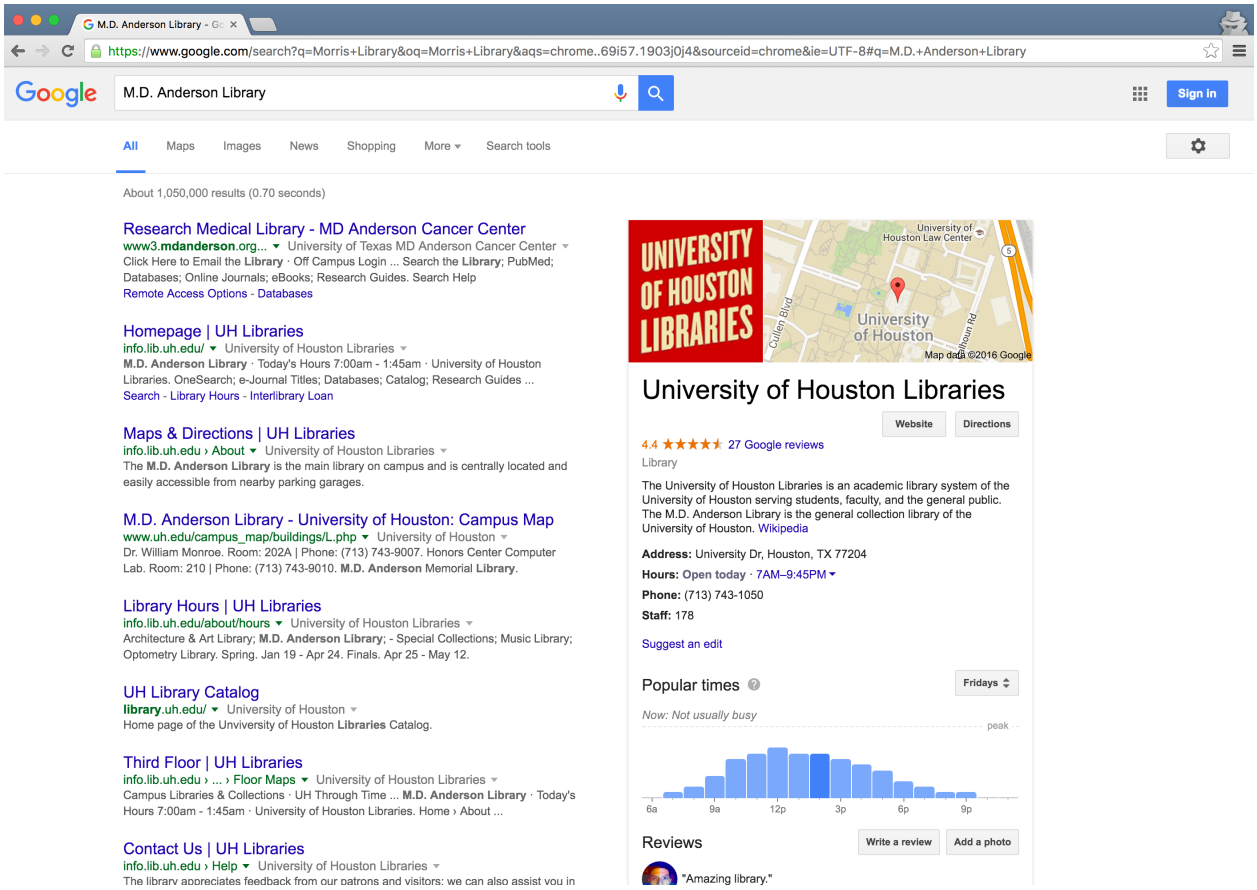

Figure 25: Search for M.D. Anderson Library shows KC for University of Houston Libraries

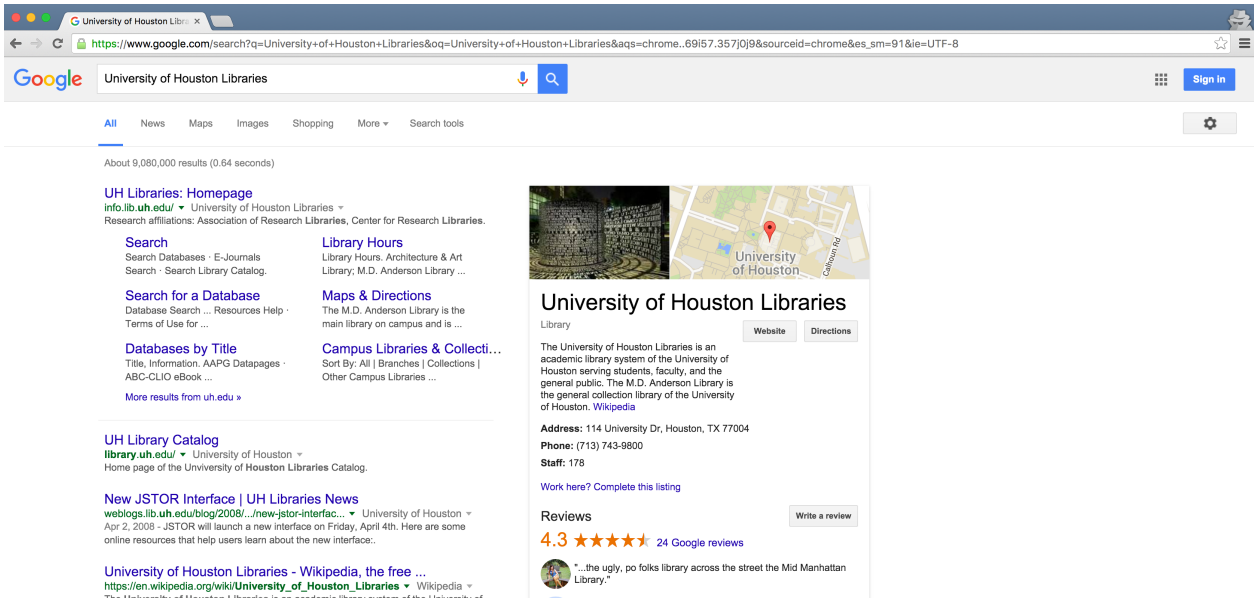

Figure 26: Search for University of Houston Libraries shows a different KC than in the previous figure.

In some cases, Google+ profiles exist for units within libraries, but not for the overall library organizations. Figure 27 shows that a search for the primary name “University of

California Berkeley Library" retrieved an unverified Google+ profile for the library's *Data Lab,* and a search for "UC Berkeley Library" retrieved an unverified Google+ profile for the library's *Government Information* department (see Figure 28). A search in Google+ for "Doe Library" failed to retrieve any result for the alternate name of the main library on campus (see Figure 29). It appears likely that enterprising employees created Google+ profiles for the *Data Lab* and the *Government Information* department, without working through the GMB process that would have generated verified Google+ profiles. A review of other screen captures confirms that businesses had not been claimed for either the UC Berkeley Library or for the Doe Memorial Library. Again, this evidence points to a lack of coordination and systematic approaches to creating and curating SWI.

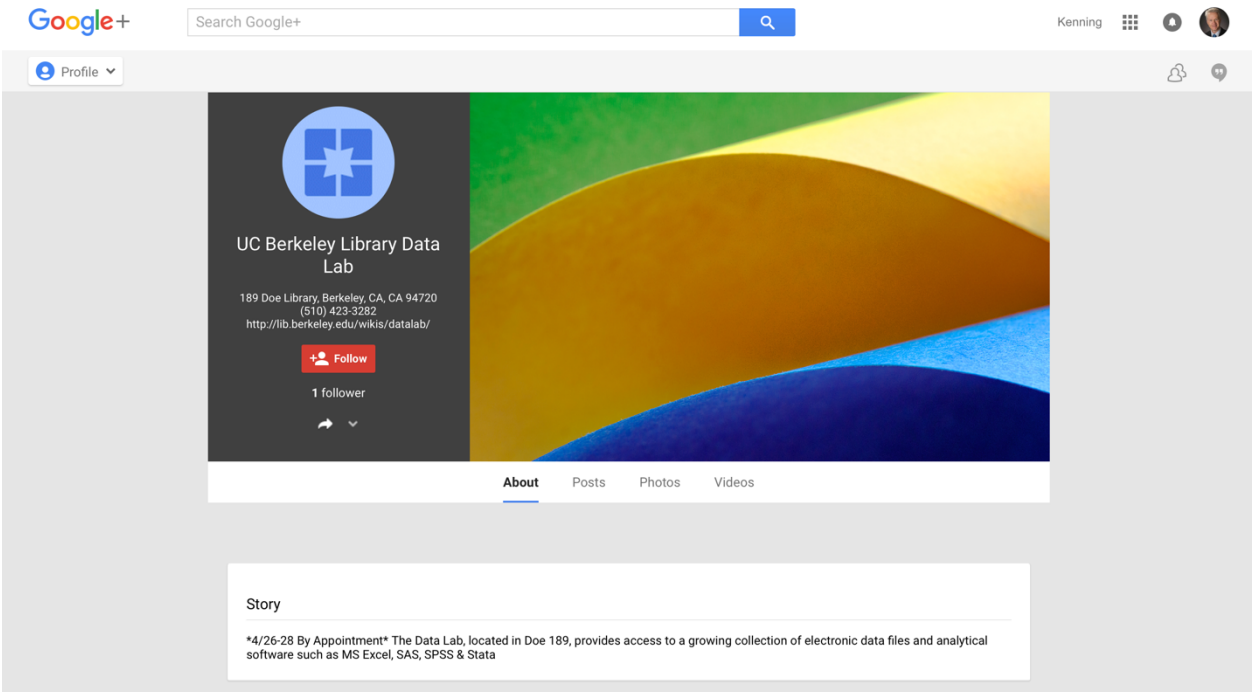

Figure 27: Search for University of California Berkeley Library retrieved a Google+ profile for the UC Berkeley Library Data Lab

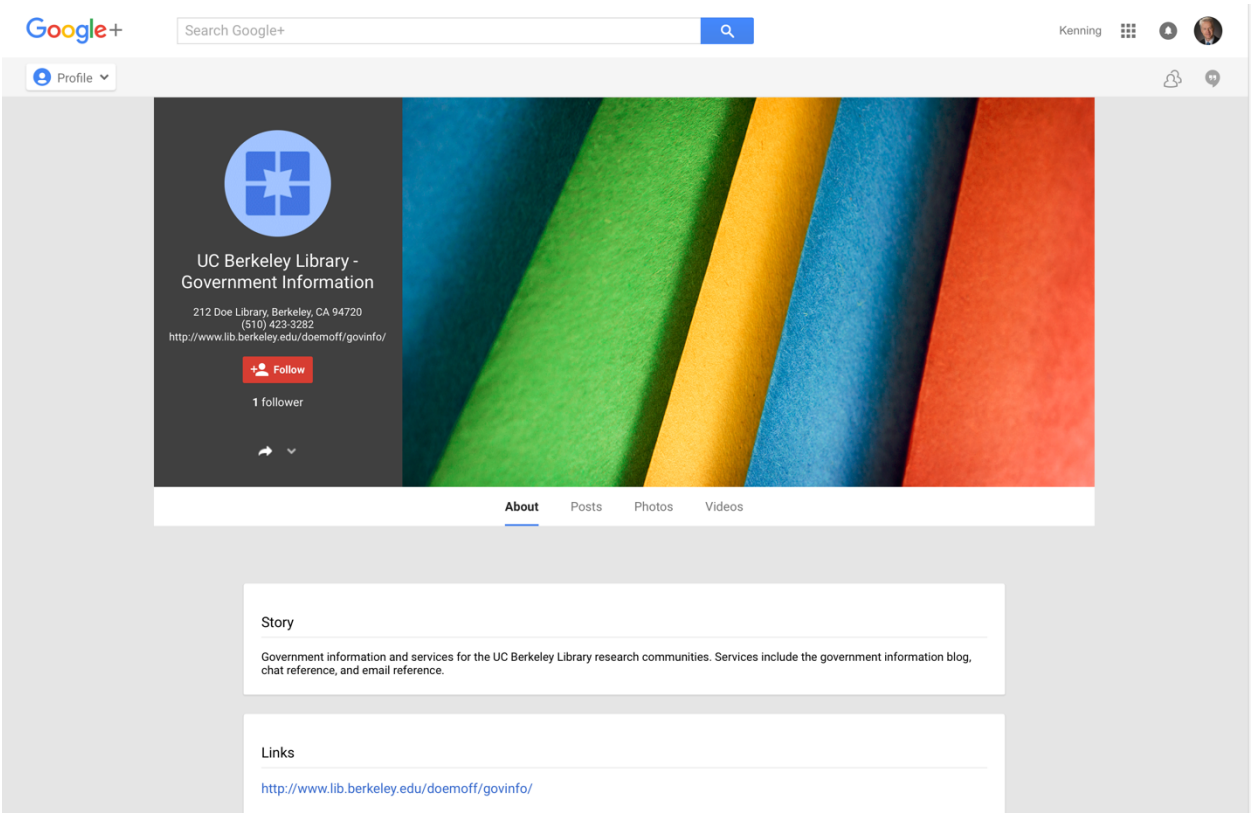

Figure 28: Unverified Google+ profile for UC Berkeley Library's Government Information Dept.

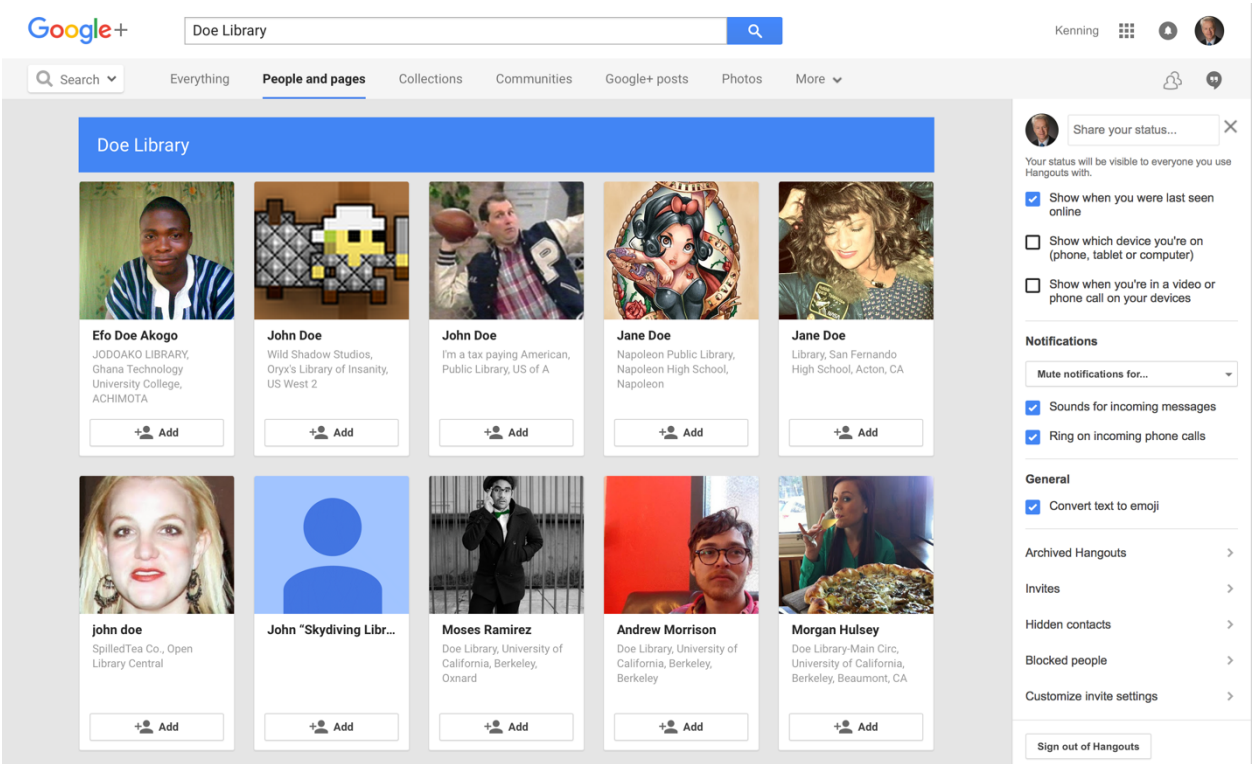

Figure 29: Google+ search for Doe Library failed to retrieve a profile.

Section 5.2.3.3 Wikipedia

ARL libraries were sometimes represented as sections in the articles for their parent institutions, but in these cases, they were not counted because the articles did not stand on their own. At the time of data collection Wikipedia contained stand-alone articles for only 82 of the 219 ARL primary or alternate library names (37%). Only 56 of those 82 Wikipedia articles included the infoboxes that are so useful to DBpedia in its generation of structured data records from Wikipedia articles. Wikipedia articles with infoboxes appeared for 30 of the 125 primary names of ARL libraries (24%) and 26 of the possible 94 alternate names (28%).

Although a full evaluation is beyond the scope of this study, it was observed that existing Wikipedia articles for ARL libraries were in various states of condition. The existing Wikipedia articles that included infoboxes were populated to a varying degree, with some showing only a few fields. The effect of this on DBpedia will be described in the next section. Articles themselves ranged enormously in length and quality, and some of the articles had been flagged by Wikipedia editors as needing additional citations to confirm the accuracy of their content, placing them under threat of deletion. Figure 30 shows an article for an ARL library that lacks an infobox and has been flagged by Wikipedia editors as needing additional citations.

Wikipedia seems to be most useful to the Google Knowledge Graph as a source of free text that is used to populate the “Description” element in the KC. Google appears to use the first sentence of the Wikipedia article verbatim for the description in the KC, so academic organizations would do well to craft that first sentence very carefully. It may be more true for Wikipedia than the other knowledge bases that ongoing and active curation of article content is required.

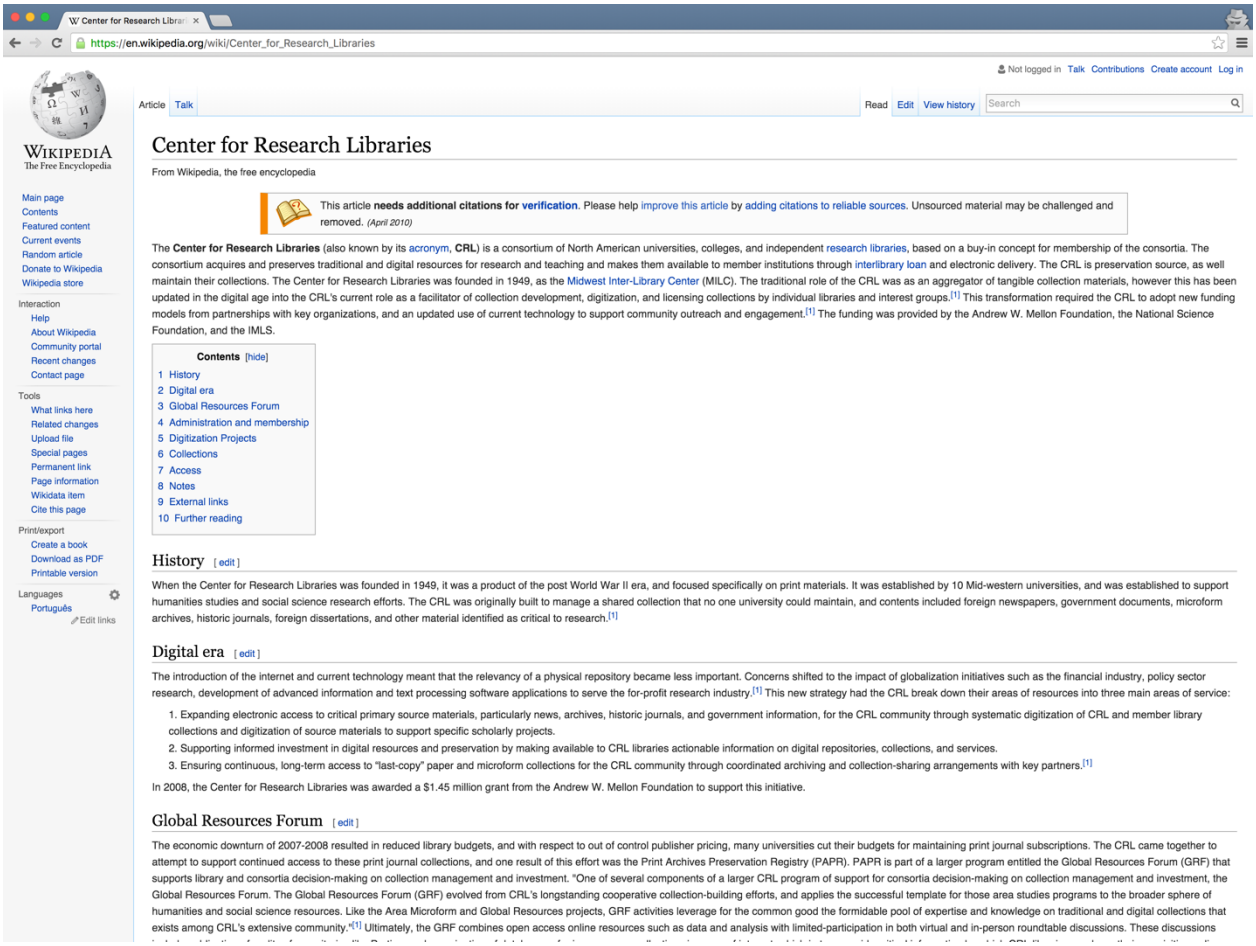

Figure 30: Wikipedia article for Center for Research Libraries, lacking an infobox and showing a flag requesting additional citations

Despite criticism that it is full of "libelous content" (Lih 2009), that its contributors are mostly male (Meyer 2013), and that its content may be biased in terms of gender (Reagle and Rhue 2011) and culture (Callahan and Herring 2011), Wikipedia has continued to grow and the English version now contains nearly 5.3 million articles (Wikimedia Foundation, Inc. 2016). In addition, Wikipedia is a formidable player in terms of traffic it receives to its website. The English-language version of Wikipedia is the sixth most visited website in the world and fully 41% of its traffic is directed there from search engines, indicating that its articles are well indexed (Alexa Internet, Inc. 2016).

As is the case for GMB and Google+, a process exists for creating and curating Wikipedia articles ("Wikipedia: Contributing to Wikipedia" 2016). An understanding of Wikipedia values and culture is paramount. The encyclopedia is known for its community of volunteer editors who are quick to delete articles that they feel were created for purposes of self-promotion rather than for informational purposes. In short, an article about an academic organization should be similar in tone to a scholarly article published in an

academic journal, and it should include citations to support factual claims. While these citations need not be online, they must be publicly verifiable and merely citing other Wikipedia articles is also not acceptable. Organizations that publish articles with citations to predominantly internal or self-published reports or documents are at risk of being removed from the encyclopedia.

Because Wikipedia is a community-managed encyclopedia, organizations should be aware that editors outside the organization will likely contribute information to the articles after they have been published. Effort can be minimized and the power of the community can be leveraged when the organization creates a seed article and allows others to add content. While most community contributions are well intentioned and useful, monitoring the articles for new additions is a good tactic to safeguard against inaccurate claims that might damage the organization's reputation.

Section 5.2.3.4 DBpedia

DBpedia is unusual in this group of knowledge bases because it does not facilitate account holder or community interaction to create or edit records. Instead, records are automatically generated from Wikipedia articles. Conceptually, DBpedia offers very useful records for the Semantic Web, since data elements are structured as linked data and are made available as SPARQL endpoints. However, despite DBpedia's acknowledged centrality in the LOD cloud and the wealth of structured data that it offers freely, the evidence gathered for this dissertation did not establish a connection between DBpedia records and the presence of KC.

While DBpedia developers intended to "regularly update the DBpedia knowledge base with the [monthly] dumps of 30 Wikipedia editions" (Bizer et al. 2009), the actual publication of new data sets has fallen to an annual schedule. Although DBpedia Live is updated almost instantaneously from Wikipedia, its records are not available as a downloadable data set. The growth of Wikipedia is the likely cause of the delay in availability of the downloadable linked data set. In 2009, DBpedia reported 2.6 million entities and 274 million RDF triples in its knowledge base (Bizer et al. 2009), and in 2016 those same metrics had jumped to 6.2 million entities and 8.8 billion triples. Perhaps as a result of this explosive growth, the data set extracted from Wikipedia in October 2015 only

became available in DBpedia in March of 2016 (Freudenberg, Kontokostas, and Hellmann 2016). This lag slowed some of the research in this dissertation, as the author was operating under the assumption that the linked data set from DBpedia was a source of information for the Google Knowledge Graph, and it was difficult to confirm that assumption while the data set remained unpublished. In the end, no connection to the Google Knowledge Graph could be confirmed.

Although "the type of wiki contents that is most valuable for the DBpedia extraction are Wikipedia infoboxes" (Bizer et al. 2009), this study shows that it is possible for a library to have a DBpedia record without an infobox in its Wikipedia article. However, it does appear that Wikipedia articles without infoboxes result in much smaller DBpedia records than Wikipedia articles with infoboxes. An example of a robust DBpedia record can be seen for the Library of Congress (see Figure 31), whose Wikipedia article includes an infobox with numerous populated fields. It should be noted that the actual DBpedia record in this example extends well below the limits of the screen that was captured.

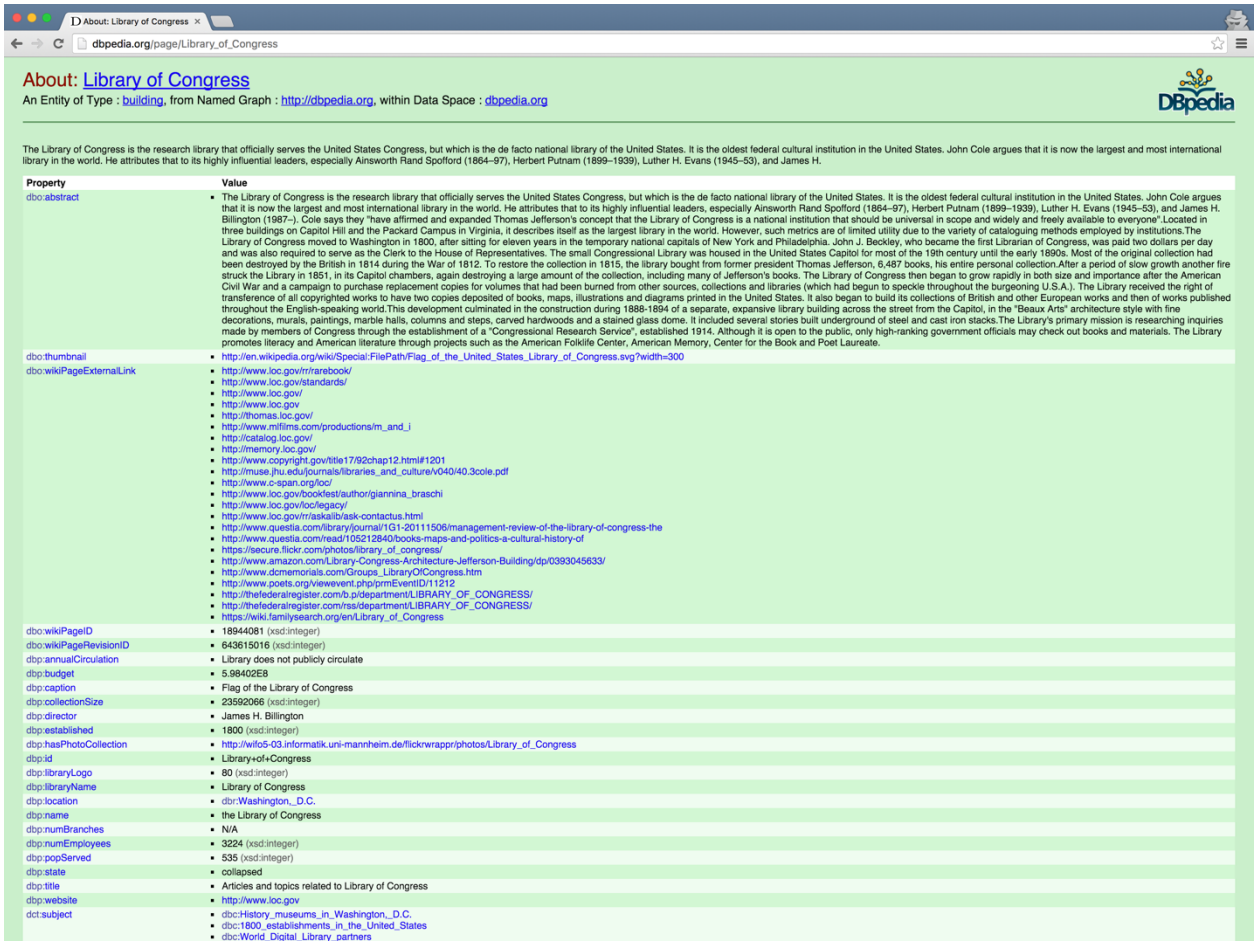

Figure 31: Library of Congress DBpedia record

Examples of sparse DBpedia records are shown, below, for Emory University (see Figure 32), Rice University (see Figure 33), and Tulane University (see Figure 34). The Wikipedia article for each of these libraries lacked an infobox.

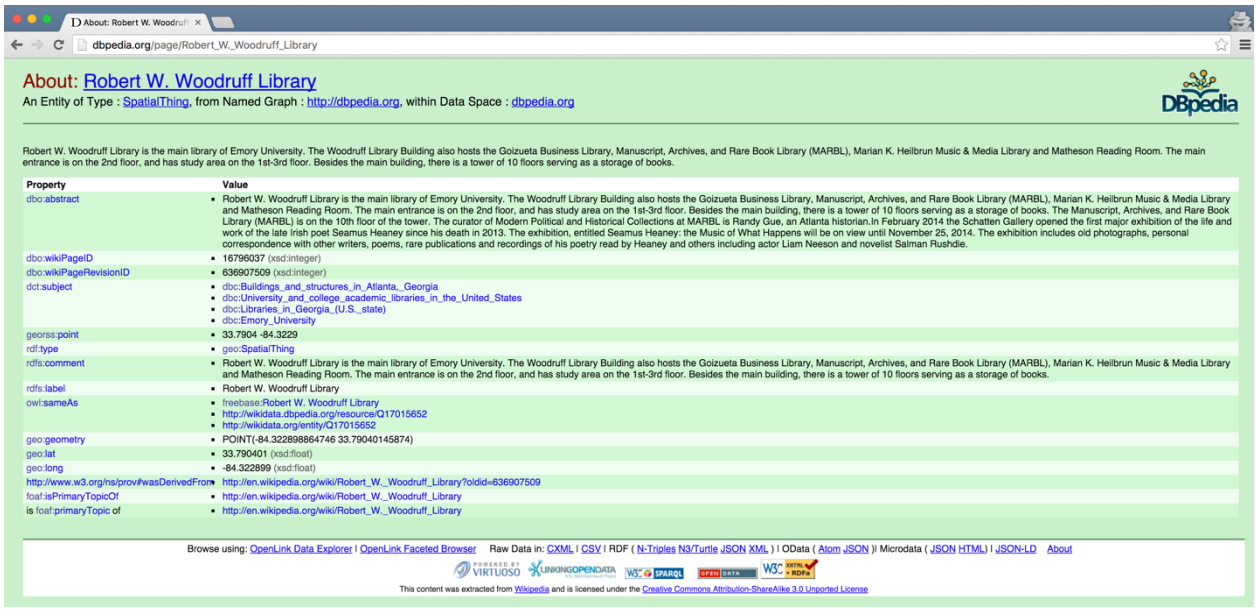

Figure 32: Minimal DBpedia record for Robert W. Woodruff Library at Emory University

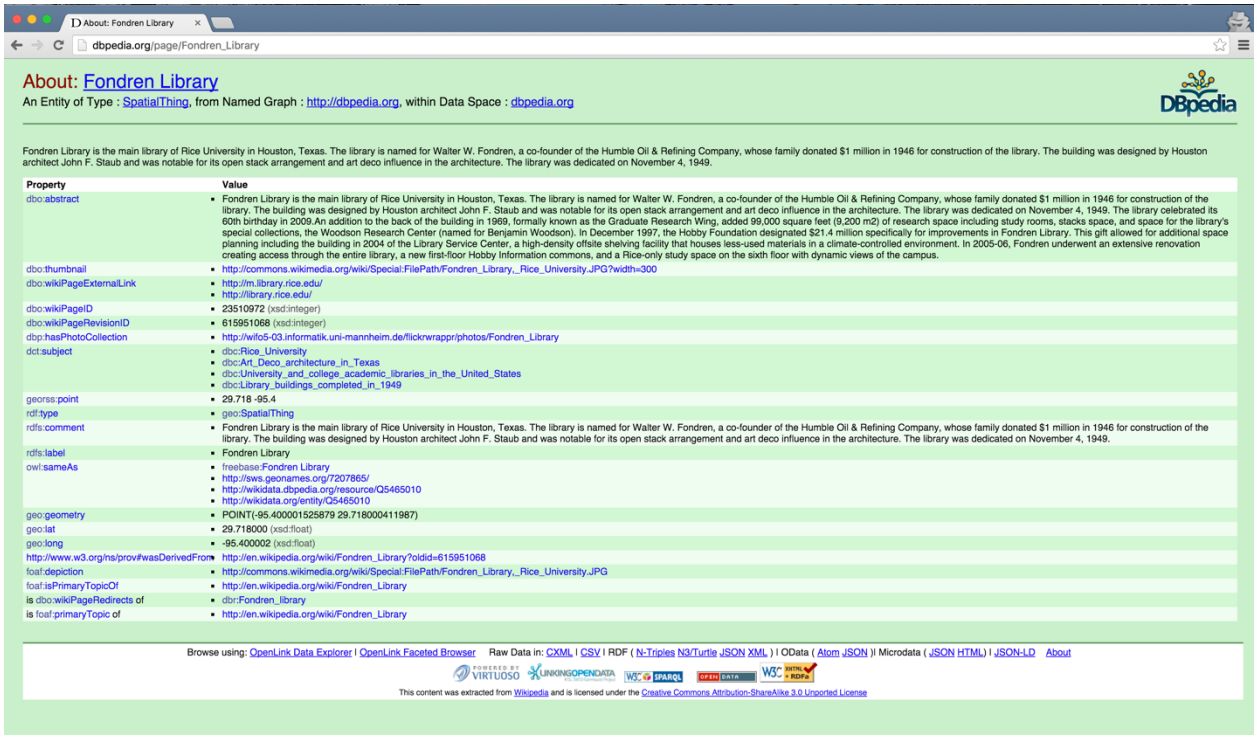

Figure 33: Minimal DBpedia record for Rice University's Fondren Library

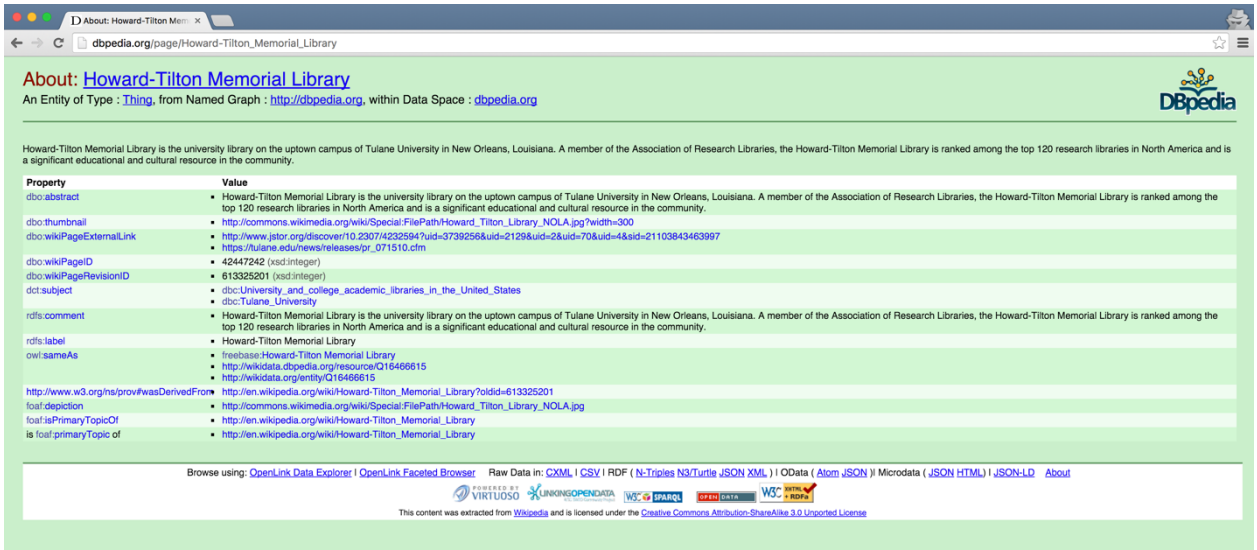

Figure 34: Minimal DBpedia record shown for Tulane University's Howard Tilton Memorial Library

Because of concerns for multicollinearity, DBpedia was not included in the logistic regression analysis that tested influence of the knowledge bases on the appearance of the information elements on the KC. It is unlikely that DBpedia plays any direct role in the appearance of KC, but this does not diminish the value of the knowledge base on the Semantic Web. It offers unparalleled structured data records in several formats that are finding use by other services, such as the BBC, which utilizes DBpedia metadata to automatically build its own controlled vocabulary (Raimond et al. 2010). Since DBpedia records are generated automatically from Wikipedia articles, without further interaction required by article authors and editors, there is additional value gained when academic organizations create and maintain Wikipedia articles for themselves.

Section 5.2.3.5 Wikidata

Wikidata is the newest of the five knowledge bases that were tested in this study, but the results of the logistic regression analysis indicate that it may already be playing a role in Google's Knowledge Graph. This makes sense since Google has dramatically increased Wikidata content by migrating its Freebase records into Wikidata (Tanon et al. 2016a). Freebase was acknowledged as a primary data source for the Google Knowledge Graph prior to its demise (Butzbach 2014). Wikidata's structured data record format should be much more useful to Google than Wikipedia articles that are mostly unstructured text, but it begs the question as to why Google doesn't seem to make use of DBpedia. Regardless, Wikidata is a knowledge base that is easily accessed and maintained, and delivers a return on SWI according to the time and energy devoted to it.

A further study into the completeness of Wikidata records and their effect on KC is warranted. Although this study eliminated Wikidata records if they contained only a single field that referenced the Wikipedia identifier, other records that were included sometimes contained only minimal geographic fields, indicating that these might also have been automatically generated. In other words, the number of Wikidata records that are being explicitly created and curated by academic libraries for their own organizations are probably lower than calculated in this study.

Section 5.2.4 Sub-question 1

This sub-question tried to determine the likelihood of an accurate KC displaying if the ARL library had not claimed and verified its business in GMB. The data reveal that 28% of primary library names that displayed an accurate KC during data collection did not show a profile in GMB, while 37% of alternate names showing an accurate KC also did not have a profile in GMB. This indicates that it is entirely possible to have a KC without having claimed and verified a business in GMB, and that Google has most likely gathered enough verified facts from other sources to generate a KC. However, since only 6 library names with a GMB profile did not display an accurate KC and 81 had neither an accurate KC or a GMB profile it is also clear that the likelihood of having a KC is very high if the business is claimed and verified in GMB. This result suggests that claiming and verifying their business in GMB is the single most effective action that libraries can take to prompt Google to generate a KC. Again, further study of this is warranted since a significant limitation of this research was that the author did not have account holder access for any of the organizations, except for the three that are described in the case study, and therefore was unable to view most of the GMB profiles. It would be interesting to learn how the completeness of a GMB record affects the generation, accuracy and robustness of a KC.

Section 5.2.5 Sub-question 2

The second sub-question tried to determine the role Wikipedia plays in determining the presence of a description field in the KC. Google references Wikipedia with a link below the descriptions when they appear on the KC, but the data collected in this study provided an opportunity for some statistical certainty as to whether a Wikipedia article contributes to the robustness of the KC. The data show that 36% of accurate KC displayed a description that could be linked to Wikipedia articles, while 45% of accurate KC lacked a description field and lacked a Wikipedia article. However, descriptions of the libraries also appeared on 10 KC when no corresponding Wikipedia article could be found, indicating that there is a small possibility that Google can gather descriptions from other places, such as GMB profiles, when no Wikipedia article is available. Overall, the presence of a Wikipedia article does tend to result in descriptions on accurate KC, as shown in Figure 12 in Chapter 4.

In some cases Google has determined "same as" relationships of primary and alternate names, and it displays the description for the library name under which the Wikipedia article exists. For instance, a Google search for *Rice University Library* displays a KC with description for the *Fondren Library* on the Rice University campus (Figure 35). A Wikipedia article for *Rice University Library* does not exist (Figure 36), but an article for *Fondren Library* does exist (Figure 37). Google has associated the *Fondren Library* with *Rice University Library* and has used the Wikipedia description for *Fondren* in the KC when "Rice University Library" is the search string. Google's ability to associate these names probably comes from a well-developed record in GMB, as searches for both library names in GMB point to the same record for the *Fondren Library*.

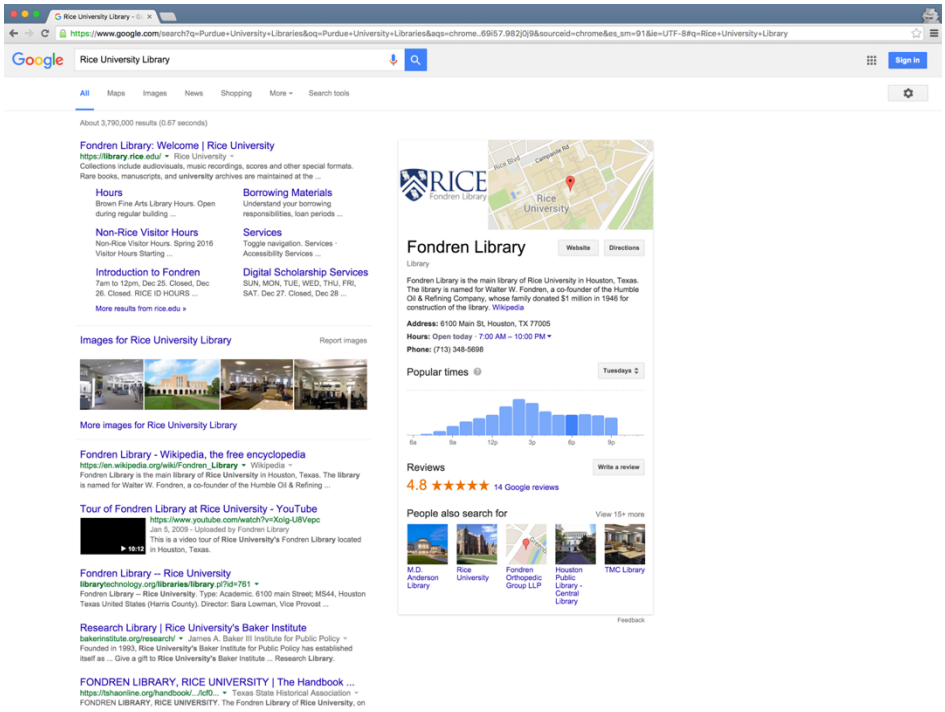

Figure 35: Google search for Rice University Library displays a KC with description field for the Fondren Library at Rice University

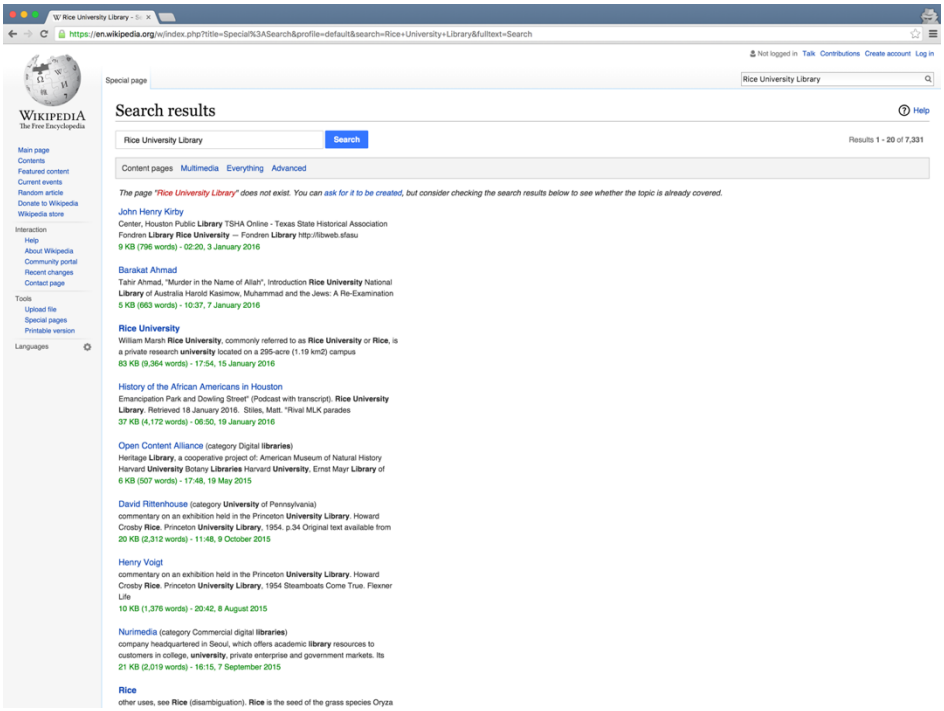

Figure 36: Screen capture showing that Wikipedia article for Rice University Library does not exist

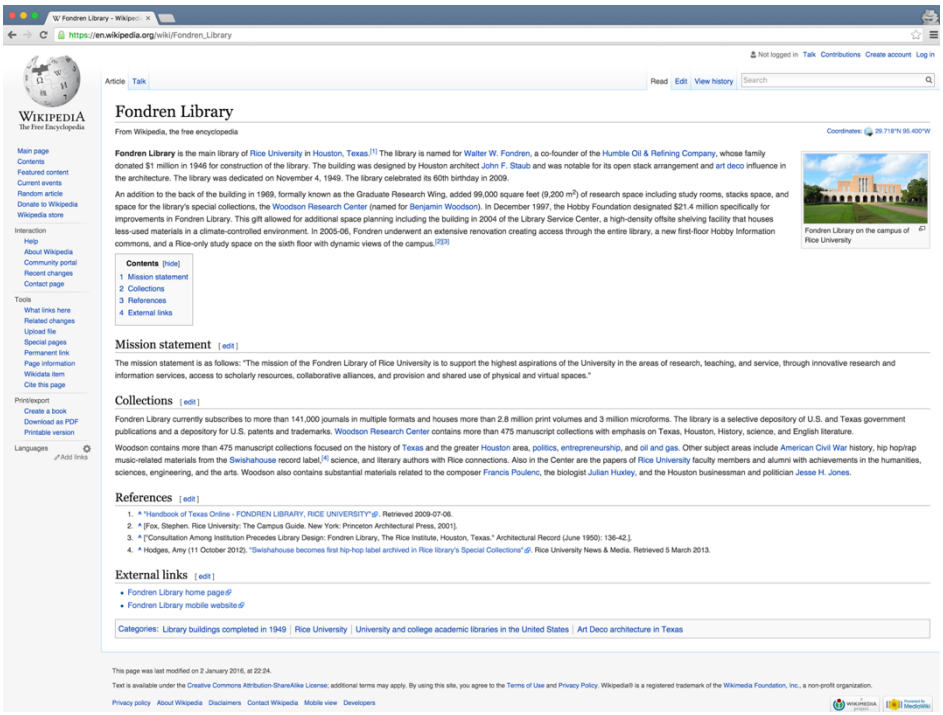

Figure 37: Screen capture showing existence of Wikipedia article for Fondren Library at Rice University

The results of the findings demonstrate that academic organizations can expect textual descriptions to appear on their KC if they have created a Wikipedia article for the organization. As ever, consistent use of the organization's name and explicitly establishing

"same as" relationships of primary and alternate names is important to the search engine's comprehension and appropriate display of KC.

Section 5.2.6 Research Question 3

The logistic regression analyses in this study estimated the odds of three types of knowledge base records (GMB, Wikipedia, Wikidata) predicting the appearance of three groups of information elements that commonly appear in KC. The first group (Description) comprised a single element, i.e., the free-text description that can appear in a KC and appears to be drawn from Wikipedia. The second group (Appearance) comprised three elements: Image; Logo; Type. The third group (Contact) comprised four elements: Address; Telephone number; Website link; Directions.

The strongest indicator of prediction appeared in the logistic regression for the Description group. Wikipedia shows the strongest odds of predicting the appearance of the description field in the KC. The odds ratio factor indicates that a description field is 5.4 times more likely to appear on the KC than if a Wikipedia article did not exist. This prediction confirms Google's indication on the KC that it draws the description for the organization from Wikipedia.

Surprising in this analysis is the strength of Wikidata as a predictor of the description field on the KC. With a factor of 4.8, Wikidata is only slightly behind Wikipedia in predicting the odds of appearance of a description filed. The reason this is surprising is that although Wikidata has the capacity in its record structure for a free text description, it was the rare ARL library that had populated this field in its Wikidata record. Furthermore, only 29% of ARL libraries had a Wikidata record for either their primary or alternate names, and some of those records were only minimally populated. The data collection period (late December 2015-April 2016) for this study coincided with the migration project that moved "14 million new statements" from Freebase to Wikidata by January 2016 (Tanon et al. 2016b), so it is possible that the migration had an effect. However, the effect of multicollinearity cannot be ruled out, either, because many Wikidata records are also initially generated from Wikipedia articles. The very small number of ARL library records that appeared in Wikidata and the potential multicollinearity with Wikipedia renders the logistic regression for Wikidata suspect.

The odds of the Appearance group showing in a KC had a prediction factor of 2.2 for GMB, but the confidence interval showed the factor lacked any statistical certainty. The small data set of only 66 GMB records for ARL primary and alternate names may have precluded any certainty. Wikipedia and Wikidata showed almost no effect on the Appearance group, but also without any statistical certainty. The situation was similar for the Contact group.

It may be impossible to predict whether any knowledge base has a much greater effect on KC robustness than any other due to trust and multicollinearity. The concept of multicollinearity was introduced in the Findings chapter to explain that Google+ and DBpedia had been removed from the logistic regression model, because they were not independent of GMB and Wikipedia, respectively. They were therefore unlikely to add independent signals that could be shown to influence (or not influence) the appearance of the information elements. It now appears that the logistic regression results for Wikidata might also be attributed to multicollinearity.

It is worth noting a parallel concept to multicollinearity. There is evidence that search engines like Google establish trust and verification by using several (possibly even numerous) data sources to confirm facts about entities. Tanon et al. provide an oblique reference to this when they discuss entity mapping, which "deals with finding the objects in several sources that refer to the same entity in the domain of discourse" (Tanon et al. 2016b). The trust and verification that Google seeks for its Knowledge Graph is similar to the trust and verification that help support the sharing economy (Ert, Fleischer, and Magen 2016; Hamari, Sjöklint, and Ukkonen 2015). Airbnb, Inc., is a prominent example in this economy, as it has developed a system it calls Verified ID, which requires both parties to a lodging transaction to provide online and offline data sources that Airbnb can verify to confirm identity (Lawler 2013; Guttentag 2015). In this case, trust and verification is not just important for assembling accurate facts, but may help avoid criminality in an industry where people welcome strangers into their homes. The multiple sources from which Google draws to establish its trust in the claims about an organization, and the proprietary nature of the Knowledge Graph, makes it difficult to predict with certainty the actions that will generate a KC. However, the overall findings of this study do suggest that the odds of appearance and robustness of a KC will increase because of investments of effort in GMB, Wikipedia, and Wikidata.

Section 5.3 Review of Case Studies

The process to establish SWI at the *MSU Library* began in early 2013 and continued development through the work with *McMaster University Library* and the *Coalition for Networked Information (CNI)*, eventually resulting in a refined strategy that prioritized engagement with GMB, Wikipedia, and Wikidata. Initially, the author and his colleagues focused on creating Wikipedia articles, as they expected the resulting DBpedia record would be utilized by Google, an expectation that the findings of this dissertation has refuted. By the time the *McMaster University Library* study was implemented it was clear that GMB played a much stronger role in KC production, and with *CNI* the process was much more advanced.

The case studies each support the importance of naming consistency and account ownership. There were at least two Google+ profiles for the MSU Library, and some naming inconsistency was evident between the *Montana State University Library* organization name and the *Renne Library* building name. In 2013, Freebase was still acknowledged as "one of the largest repositories of data Google uses to construct the Knowledge Graph" (Butzbach 2014), and it appears that Google was unable to reconcile the organization formally known as the *Montana State University Library* with its *Renne Library* Freebase record. This condition may have contributed to the KC being displayed for the branch library in Billings, MT rather than Bozeman, in 2012.

When the SWI process was further developed at *McMaster University Library* in 2015, Google had already announced the planned shutdown of Freebase, so there was no point in expending time or energy in that knowledge base. However, McMaster had a more serious problem with name consistency and account ownership. The SWI of the umbrella organization known as *McMaster University Library* was competing with the brands of its smaller campus libraries (*Thode* and *Innis*) as well as with its own building name: *Mills Memorial Library*. Difficulty in confirming the street address also added to its problems. Through patient intervention with GMB by the McMaster University Library AUL, several of the records were eventually merged and the unverified Google+ profiles were deleted. Today, a search for *McMaster University Library* displays an accurate and robust KC, with a description drawn from the Wikipedia article that was published in 2016.

The CNI case study also required significant intervention to align or delete several accounts and properties that had been created over the years for various reasons. Despite this behind-the-scenes management, the KC did not begin to appear in a search for *Coalition for Networked* Information until the acronym "CNI" was prepended to the name of the organization in the GMB record to match the name on the organization's website, i.e. *CNI: Coalition for Networked Information*. The Wikipedia article name was also edited to match. The acronym "CNI" unfortunately competes with the stock ticker symbol for the Canadian National Railway, so it is unlikely that a Google search for "CNI" will ever display the *Coalition for Networked Information* at the top of the SERP along with its KC.

Ultimately, the SWI process resulted in KC for all three organizations. The success of these three case studies has helped form the SWI services at the MSU Library, which will be explained in the next chapter.

Section 5.4 Other Factors of Interest

The impact of academic libraries' inconsistent use of their primary and alternate names on the Semantic Web was a surprising and notable finding of this research. Variable use of names in the analog environment is less problematic because of human ability to elucidate context and relationships. In the machine-based digital environment of the Semantic Web, however, use of different names in different contexts without explicitly creating "same as" relationships for machine comprehension, can severely impact SWI. A more detailed description of the origin and problems of primary and alternate names follows.

Section 5.4.1 Primary Versus Alternate Names of Organizations

Many academic organizations and the buildings in which they reside have more than one name. Alternate names are most often the result of financial gifts to the university, in recognition of which the organization or building has been bestowed with the donor's name. The name that organizations use to refer to themselves may vary based on the situation. There is nothing inherently right or wrong about the use of these primary or alternate names, but their inconsistent use on the Semantic Web sends mixed signals to machines that often have no frame of reference from which to match the same organization to different names. The data collected in this study have demonstrated that ARL libraries

use their name variations inconsistently across their websites and across the knowledge bases that help populate Google's Knowledge Graph. This can confuse people, but it is more likely to cause confusion in the machine-based environment of the Semantic Web. Few people outside the University of Rochester are likely to know that its main library is known locally as the *Rush Rhees Library*, or that *Firestone Memorial Library* is the main library at Princeton University. Internet users unfamiliar with these universities are much more likely to search for them by their primary names: *University of Rochester Libraries* and *Princeton University Library*. If the machines that are interpreting search queries have no information in their databases linking both names to the same organizational entity then they will be less likely to return an accurate KC, let alone refer users to the correct website.

In the case of the member libraries of the ARL, each organization has approved the listing of its primary name in the ARL directory (see Appendix A), and one would presume this is the official name by which the organization would like to be known. However, only 46% of the primary names searched in this study displayed an accurate KC. Google+ profiles appear for both primary and alternate names, as well-intentioned staff members have sometimes created profiles for departments within libraries, even when no profile in Google+ exists for the library itself. The problem of inconsistent use of names extends to the other knowledge bases in this study, and thus, the lack of comprehension is perpetuated across the Semantic Web.

The University of Washington is a particularly complicated case, although it is by no means the only one. Listed as *University of Washington Libraries* in the ARL directory, the separately named but physically conjoined *Suzallo and Allen Libraries* are considered the main library on the flagship campus. Although a website with the name *University of Washington Libraries* exists, a Google search for that name displays a KC for the *Social Work Library* on the UW campus (see Figure 38). A separate website for the *Suzallo and Allen Libraries* does exist, but there is no KC for that entity. Knowledge base records have presumably been created at different times by different people, as the names used in those records are not consistent. Table 8 shows name variations for the main library at UW.

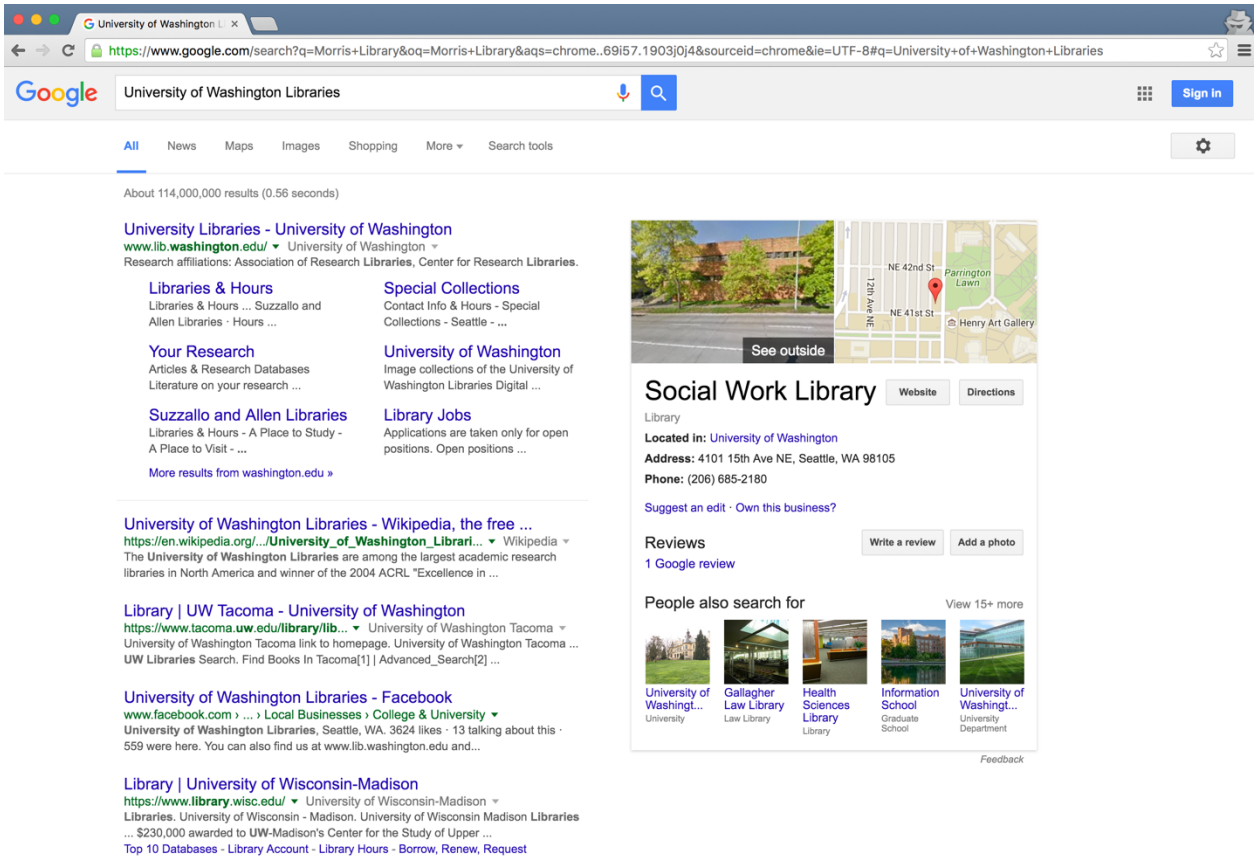

Figure 38: Google search for University of Washington Libraries displays KC for the Social Work Library

Library Name	Website	KC	Google My Business	Google+	Wikipedia	DBpedia	Wikidata
University of Washington Libraries	Yes	No	No	Yes, verified	Yes, flagged, no infobox	Yes, limited	No
Suzallo and Allen Libraries	Yes	No	No	No	No	No	No
Allen Library	No	Yes	Yes	Yes, verified	No	No	No
Suzallo Library	No	No	No	No	Yes	Yes	Yes

Table 8: Name variations and results for main University of Washington libraries

The University of Washington example is not unusual and is indicative of a lack of consistent communication practices regarding Semantic Web representation of organizations on university campuses. The problem should be addressed by a well-formed

communication plan that is driven by the organization's administrators, but although we learned earlier that marketing and outreach was prominent in 71% of academic library strategic plans, the concepts of naming consistency, let alone feeding knowledge bases on the Semantic Web, were absent from the article that reviewed those plans (Saunders 2015).

Section 5.4.2 Physical Addresses of Organizations

Physical addresses of organizations are crucial for semantically enriched applications that are widely used on mobile devices. Mapping applications cannot determine the correct location, calculate distance, or offer specific directions if the physical address registered in knowledge bases like GMB is missing or inaccurate. Structural inconsistencies exist on university campuses that make this a problematic issue for machine comprehension, because many university buildings don't have individual street addresses. Instead, they have post office (P.O.) box numbers and mail for the organizations is typically sorted and delivered by a central facility. While this central mail drop system is efficient, it complicates GMB's verification process because GMB doesn't accept P.O. box numbers as mailing addresses, and the result is the postcard that Google mails often doesn't reach the claimant. These factors can make it difficult for organizations to complete the GMB verification process.

MSU Library serves to illustrate the problem. The organization officially called the *Montana State University Library* is located in a building known as the *Roland R. Renne Library*, named after a former president of the university. Consequently, many people in the campus community refer to the library organization as the *Renne Library*, but this is inaccurate because "Renne" is the name of the building and not the organization. To complicate matters further, the *Renne Library* building has long housed the campus Information Technology Center, a separate organization that reports to the university's Chief Information Officer. Each organization has a different P.O. box number but no universally recognized mailing address. The GMB verification process for the MSU Library was hampered because the postcard that Google mailed for verification was lost. Through some direct negotiation the library was eventually able to verify its business claim with GMB, but without a separate mailing address for the IT Center GMB would have difficulty verifying the physical location of the IT Center after the Library had already claimed the building as its address. Physical addresses of organizations in universities, then, are not

conducive to providing the data needed for Semantic Web knowledge bases and the technology applications that use them. This is another example of the complicated transition from the analog world in which people have long worked, to the digital world that is increasingly managed by machines.

Section 5.5 Summary of Discussion

This chapter offered in-depth analysis and speculation about the findings from the data set. Generally, the findings have demonstrated that the condition of SWI, as indicated by the presence of accurate KC, is poor when the primary names of most ARL member libraries were searched. Alternate names fared better, but only 37% of libraries showed the same KC when their primary and alternate names were searched in Google, indicating that most of the time the search engine doesn't understand that both names are related to the same organization. The inconsistent use of primary and alternate names in library websites and knowledge base records contribute to this problem and illustrates a disconnect in the way ARL libraries represent themselves in the official ARL membership directory versus the Semantic Web.

The logistic regression analysis confirmed that Wikipedia has the greatest influence on the appearance of a description field in the KC. This finding shows that the creation of a Wikipedia article about the organization is worthwhile, and that particular care should be given to the opening sentences of the article as that is the text that usually appeared in the limited space of the KC description field. Surprisingly, Wikidata also showed influence on the appearance of a description field in the logistic regression analysis, but this finding is suspect and is likely due to multicollinearity with Wikipedia. Many Wikidata records have been automatically generated from Wikipedia, but the Wikidata records found for ARL libraries tend to be poorly populated and the clear majority lack a description field. Therefore, even though the logistic regression showed Wikidata exerting some level of influence on KC description fields, it is unlikely that the Google Knowledge Graph could have acquired descriptive text about the ARL libraries from Wikidata.

The presence of GMB records for ARL libraries that have accurate KC suggests that a KC is much more likely to appear if the business has been claimed and verified in GMB. However, the logistic regression was unable to demonstrate with any statistical certainty that GMB is a strong influence on the appearance of the KC information elements that were

grouped as Appearance and Contact variables. Since only 30% of library "businesses" had been claimed and verified for the 219 ARL primary and alternate library names, it is possible that the sample set of these positive influences was simply too small in the face of the neutral or negative influences. In other words, since only 68 of the 219 possible names had been verified in GMB, the influence that GMB records could have shown for Appearance and Contact variables was necessarily diminished by the 151 remaining names that had not been claimed and verified in GMB.

Few libraries have apparently engaged with the knowledge bases examined in this study, and it is likely that this lack of engagement has also handicapped Google's ability to establish verified facts about the organizations. Several examples of profiles and records from the knowledge bases illustrate the inconsistent and often confused manner in which the libraries represent themselves in different venues. In some cases, libraries present different names, addresses, and even telephone numbers, depending on the knowledge base. In some cases, profiles or records have been created for units within the libraries, even when a profile or record for the overall library organization does not exist. In cases where Wikipedia articles fail to include infoboxes, the resulting DBpedia record is noticeably less robust. Poor DBpedia records, while apparently not a direct influence on SWI, could have downstream effects in other knowledge bases or services that depend on DBpedia for freely available structured data records.

This chapter also briefly reviewed the three case studies that demonstrated the success of an evolving process to establish and improve SWI. Finally, it offered a discussion of the impact of primary and alternate names of the libraries, and of the physical and mailing address infrastructure of many academic institutions, which can create problems for the GMB verification process.

The next chapter is intended to briefly illustrate that lack of SWI extends beyond libraries and into other academic organizations. It also describes the services that are being implemented at Montana State University to establish and improve SWI of campus organizational units, and which other libraries could adapt for their purposes.

Chapter 6 Broader Implication of the Research

Section 6.1 Introduction

The study described in this dissertation demonstrated that most member organizations of the Association of Research Libraries are poorly represented on the Semantic Web, in terms of the KC that are displayed for them in SERP and in the knowledge bases that help populate search engine knowledge graphs. This chapter provides an example to illustrate that the problem of SWI is not limited to libraries, and in fact affects organizations throughout academic institutions. It begins with a brief portrait of SWI across the higher-level organizations of Montana State University (MSU), and then describes a set of services that are currently being designed and implemented at MSU to improve SWI for its colleges. These services could easily be adapted at other institutions, allowing academic libraries to demonstrate expertise and leadership in Semantic Web development.

Section 6.2 MSU Academic Organizations

According to The Carnegie Classification of Institutions of Higher Education there are 4,664 postsecondary degree-granting institutions that enroll over 20 million students in the United States (Center for Postsecondary Research 2016). Approximately 335 of these institutions grant doctoral degrees and are considered research universities. Most of these research universities follow a common organizational hierarchy. The top level is the institution itself, e.g. Stanford University; Cornell University, University of Wisconsin, etc. The next level within the institution is represented by the colleges; e.g. College of Agriculture, College of Engineering; College of Arts. Within each college the hierarchy continues with multiple departments, and within departments there may be research centers or institutes.

Random searches in Google by the author over the past two years have shown a pattern of SWI for research universities. A robust KC tends to display in Google SERP for institution names, but once one begins to search for names of colleges within those institutions, or for departments, centers and institutes, many fewer KC appear and those that do appear are often inaccurate or not very robust.

MSU is a mid-sized research university according to the Carnegie definition and is therefore representative of a typical U.S. research university. While many other research

universities could have been selected to help illustrate the condition of SWI beyond libraries, MSU offered a convenient and familiar sample because it is the author's home institution. The convenience extends to the SWI service that the author and his colleagues have launched at MSU, which will be described later in this chapter. Table 9 shows an SWI snapshot of eleven high-level academic organizations (colleges) at MSU in December 2015. Data for these organizations were gathered in a more limited fashion than for the ARL members. As with the larger ARL data set, the author collected screen capture files for the Google search results for each MSU organization (see Appendix D), as well as for the same five knowledge bases used for the ARL libraries. However, information elements for each KC were not measured, data were not recorded in a spreadsheet, and no statistical analysis was run for this very small data set.

College	**KC**	**GMB**	**Google+**	**Wikipedia**	**DBpedia**	**Wikidata**
Art & Architecture	No	No	No	No	No	No
Agriculture	No	No	No	No	No	No
Letters & Science	Yes	Yes	No	No	No	No
*Business**	Yes	Yes	Yes verified	No	No	No
Engineering	No	No	Yes unverified	No	No	No
Nursing	Yes	Yes	No	No	No	No
*Education, Health and Human Development**	Yes	Yes	Yes	No	No	No
Gallatin (vocational within MSU)	No	No	No	No	No	No
Graduate School	No	No	No	No	No	No
*Honors**	No	Yes	Yes	No	No	No
*Library***	Yes	Yes	Yes	Yes	Yes	Yes

Table 9: SWI of MSU colleges in December 2015

*These organizations were in the first cohort at MSU that was being served by the Library's experimental SWI service.

**The Library had successfully established its SWI in 2013-14.

The snapshot shows that only five of the eleven colleges displayed KC when searched in Google, and four of those five had either been the beneficiaries of significant intervention (e.g. the Library had been working on its own SWI since 2013), or of some lesser level of

intervention in the form of SWI services that the Library had begun to deploy across campus in late 2015. In early searches for "Montana State University College of Engineering" a KC for the Michigan State University College of Engineering would appear. The presence of articles or records in the other knowledge bases was dismal. None of the colleges (except for the Library) showed articles or records in Wikipedia, DBpedia, or Wikidata. Only half of the colleges had claimed and verified their businesses in GMB, and the only colleges that had verified Google+ profiles (EHHD, Honors, Library) had already benefited from some work by the Library. The situation at MSU reflects what the author has seen in his random searches for other research universities.

Section 6.3 Semantic Web Identity Library Services

Academic libraries are in a unique position to help other organizations on their campuses establish and maintain their SWI by offering a service that leverages traditional strengths in bibliographic management and research support. Libraries have built their reputations in part by creating and maintaining structured data records for their collections, including these well-known examples: Machine Readable Cataloging (MARC) for physical collections; Encoded Archival Description (EAD) for finding aids in archival collections; Dublin Core for digitized cultural heritage material; and the Text Encoding Initiative (TEI) for machine-readable text markup in digitized books. Establishing and maintaining SWI for academic organizations also requires engagement with structured data records, but in the environment of the Semantic Web. Cataloging and metadata librarians have an opportunity to extend their skills to offer a service that will be valued on campus, and engaging with the structured data records in Wikidata and GMB is a natural fit for these librarians.

As part of the Wikimedia Foundation's suite of products, Wikidata is community-based, meaning that anyone can create and edit records, but since shell Wikidata records are auto-generated from Wikipedia articles, it may be more appropriate to publish the Wikipedia article before trying to create and populate a Wikidata record. Like Wikipedia, the records are subject to scrutiny by a large community of editors, and well-documented support of changes is an expected part of the process. Access to GMB is more difficult and requires a Google Account, but it's possible for librarians to gain mediated access to the profiles of other organizations on campus by working closely with designated contacts.

Education and outreach are also crucial to this process, along with documentation of the process and of the steps taken.

Not all the work involves creating structured records, of course, and therefore not all the work need fall to cataloging and metadata professionals. Wikipedia articles require writing skill and must largely be developed by the people who are most familiar with the organizations in question. Library liaisons can facilitate writing the articles, ensuring that they are scholarly in tone, and that adequate research is conducted and statements are factual and supported with citations. Above all, Wikipedia articles must avoid the perception of self-promotion, otherwise they may be flagged or deleted by other editors, creating delay and even blockages. However, a large investment of effort is not required; as mentioned in Chapter 5 it is most worthwhile to create minimal articles that provide a foundation to which others can add, thereby leveraging the power of the Wikipedia editing community. It is also important to include infoboxes in the articles, but templates are available and should be used ("Category: Infobox Templates" 2016), not only because they make the process easier, but because they help DBpedia developers extract structured data for richer DBpedia records.

Applying structured data markup directly to organization websites using the Schema.org vocabulary was out of scope of this dissertation, but it is another accepted method for providing information to search engine crawlers and the graph databases they are building. Schema.org markup can help express local relationships as well as describe the architecture of the website by using the "isPartof" attribute. Schema.org can also provide additional machine understanding of entities by pointing toward DBpedia records, using the following expression in HTML markup:

---

property="additionalType" resource="http://dbpedia.org/page/…"

References of this type are not particularly useful to humans; instead they are intended to help machines gain a greater understanding of the terms and concepts used in websites. DBpedia records provide rich structured data one could expect search engines to find useful, although specific evidence of this awaits confirmation.

Section 6.3.1 SWI Services at Montana State University

In 2015 the author and several colleagues submitted a successful proposal (Arlitsch et al. 2015) to the provost at MSU, requesting two years of funding to hire a Semantic Web Identity Researcher. The two main tasks of the researcher are:

1. Establish or improve the SWI of the colleges at MSU (as well as other academic organizations as time allows).
2. Integrate the new SWI service into the workflow of the library's Resource Description and Metadata Services (RDMS) department so that it becomes an ongoing service to organizations on campus.

Section 6.3.1.1 Strategy

A coherent strategy is crucial to developing and maintaining SWI that presents academic institutions accurately and consistently on the Semantic Web. The first step to such a strategy is to raise awareness that there is a problem and that lack of attention can potentially affect the reputation and limit the performance of the entire institution. The approach that raised awareness of the issue at Montana State University included campus level conversations that were catalyzed by a discussion at a Research Council meeting in January 2015. The Vice President for Research and Economic Development (VPRED) had expressed concern that a grant proposal she submitted to the National Institutes of Health had recently been rejected, in part because the grant reviewers maintained that MSU was not engaged in the research topic in question and was therefore ill-equipped to carry out the proposed work. The VPRED pointed out that MSU was indeed engaged in this research, but for some reason the grant reviewers had been unable to ascertain that fact. While it is impossible to establish a concrete connection between that failure and poor SWI, the conversation represented a watershed moment for the author. The SWI research that he and his colleagues were conducting had been limited to libraries, but the Research Council conversation broadened the scope of the research to other campus organizations. The author realized that another potential ramification of poor SWI is that search engines might fail to connect grant reviewers to a given institution if the search engine were unable to determine that specific research occurs there. Funding agencies, relying on the recommendations of teams of reviewers, might withhold funding for grant proposals if those reviewers are unable to easily find evidence that the university has expertise or a

track record in that area of research. Additional effects were not difficult to imagine: students looking for universities that matched their research interests might never discover a program at a university that had not addressed its SWI; and researchers seeking colleagues might not connect if search engines failed to realize that their research is similar.

The author brought the topic of SWI to the provost and the other deans through a series of conversations and presentations. The presentations extended to the CIO, vice presidents and the university president, as well as chancellors and CEOs of the three other MSU campuses. One example of poor upkeep of Semantic Web data sources that the author showed these audiences was the Freebase record for MSU itself. Nearly five years after hiring its first female president of the university, the Freebase record still listed the previous president as current, and all but two of the entire university's administrative team listed in the record had since departed. The Freebase record for MSU was hopelessly out of date because no one from the university was aware of its impact and nobody was assigned to maintain it. At that point Freebase was still a significant source of information for Google's Knowledge Graph. Demonstrating the connection between the knowledge base, Google's Knowledge Graph, and the identifiability of MSU online effectively argued for the creation of an SWI strategy for MSU. The proposal to hire a Semantic Web Identity Researcher was approved and the library began to help academic organizations around campus improve their SWI.

Section 6.3.1.2 Tactics

The Semantic Web Identity Researcher (SWIR) began his work in August 2015, and less than a year into his contract he had successfully created or improved the SWI for several organizations on campus, including the *Jake Jabs College of Business and Entrepreneurship*; the *Honors College*; the *College of Education, Health and Human Development*; Campus Planning, Design and Construction; and the *Office of the Provost*. Each of these organizations now displays a KC with accurate contact information, although the robustness of some KC is limited as some of the knowledge bases are still being populated. The remaining colleges (Agriculture; Arts and Architecture; Engineering, Gallatin College, Graduate School, Letters and Science; Nursing) are participating in the second phase of SWI improvement, which is underway at the time of this writing in the autumn of 2016.

Eventually the work will move down the hierarchy to departments within colleges, and to research centers and institutes. As word of this work has spread, other campus organizations have begun to approach the library to ask for its help.

The SWIR begins his work process with each campus organization by establishing a relationship with a designated point person. Often this person has a marketing or communications role and acts as a conduit to others in the organization who might have more historical knowledge that can be utilized during the creation of records and articles in the knowledge bases. The general approach is educational, helping the point person understand the larger context of SWI and then gathering information about the organization and beginning the process to engage with GMB, Wikipedia, and Wikidata. The SWIR must continually issue cautionary words to temper the enthusiasm that a little knowledge of this subject can ignite; experience has shown that it is all too easy to make mistakes in populating knowledge bases, and some mistakes can take months to correct.

A crucial part of the initial phase is to establish baseline metrics so that progress can be measured. These metrics include the screen capture method of evidence, as described above. Additional visitation metrics are collected with the Google Analytics (GA) service, which requires embedding the JavaScript beacon code into each HTML page of the organization's website, a process that is easy in a template-driven Content Management System. GA uses this code to log visitation in its analytics software. Used in conjunction with Google Search Console (formerly known as Webmaster Tools), the service can generate reports that measure visits, show trends in user traffic, and help diagnose problems experienced by search engine crawlers as they try to harvest and index the web pages. These aspects of the process are rooted in SEO techniques.

The process of establishing SWI is necessarily collaborative, because it requires the technical expertise of the SWIR as well as the content knowledge of the point person. The process can take many months, particularly when there are delays caused by the verification process in GMB, and the protracted timeline can test the patience of all parties. Throughout this period the SWIR maintains communication with the various contact people in different organizations to foster and maintain interest. Once the KC begins to appear in search results, visitation metrics are collected for presentation to the organization's leadership team. Metrics collected thus far show an increase in website visitation, phone calls, and clicks on the driving directions link in the KC.

Section 6.4 SWI Service Example

The Honors College at MSU was one of four organizations that agreed to participate in the Library's first phase of establishing and maintaining SWI. Prior to treatment, the Honors College showed no KC, and had no presence in any of the five knowledge bases. Discussions began with Honors College personnel in September 2015, and a business was immediately claimed in GMB. The verification process took several months, due to previously discussed problems with delivery of the confirmation postcard. In December 2015, a search in Google still showed no KC (see Figure 39).

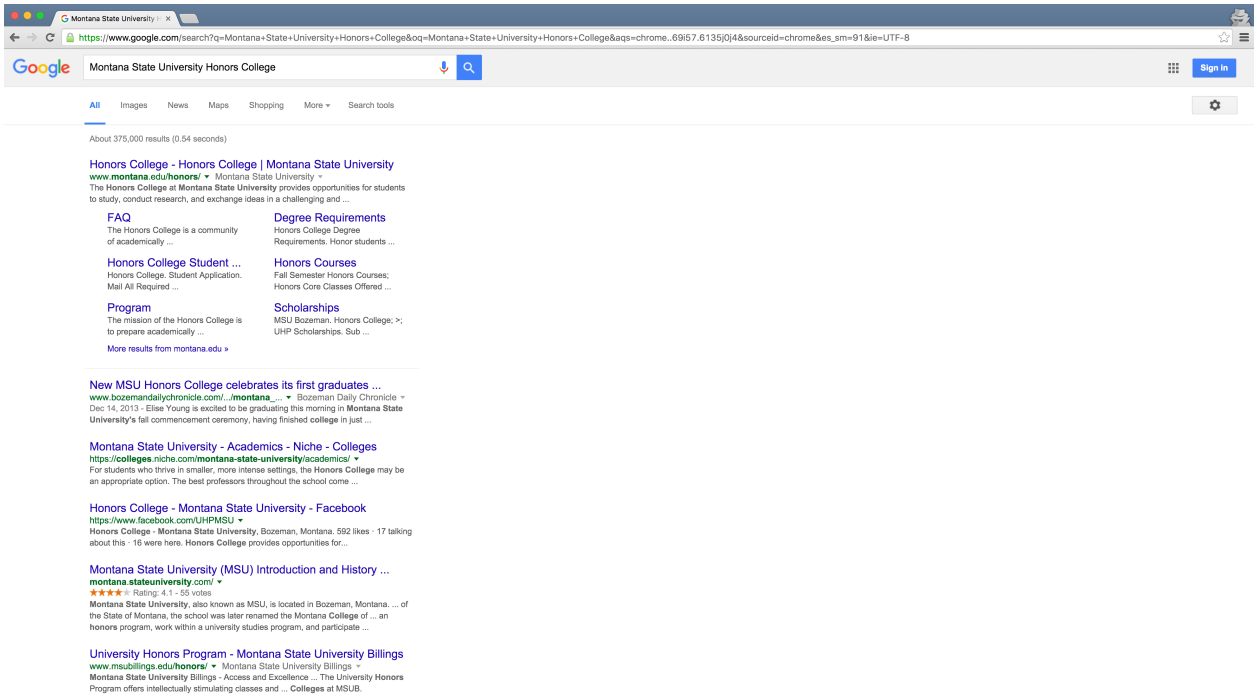

Figure 39: Google SERP for MSU Honors College in December 2015 still lacks a KC.

Creating and publishing a Wikipedia article about the Honors College also took time, but on July 11, 2016 a well-written article with an infobox was published (see Figure 40).

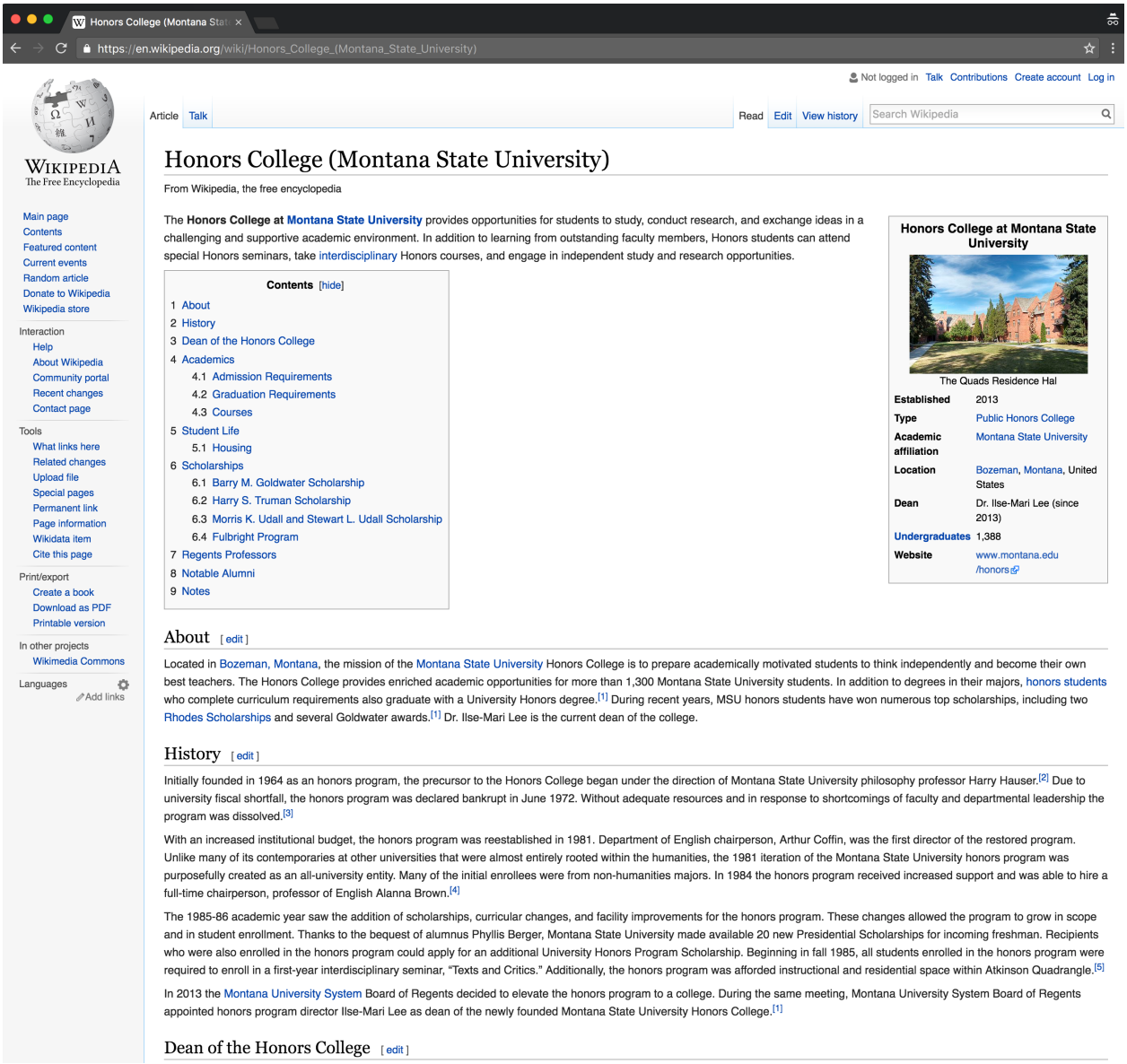

Figure 40: Portion of Wikipedia article for MSU Honors College in November 2016..

Publishing the Wikipedia article generated a shell Wikidata article, and additional fields were then filled in by the SWIR (see Figure 41)

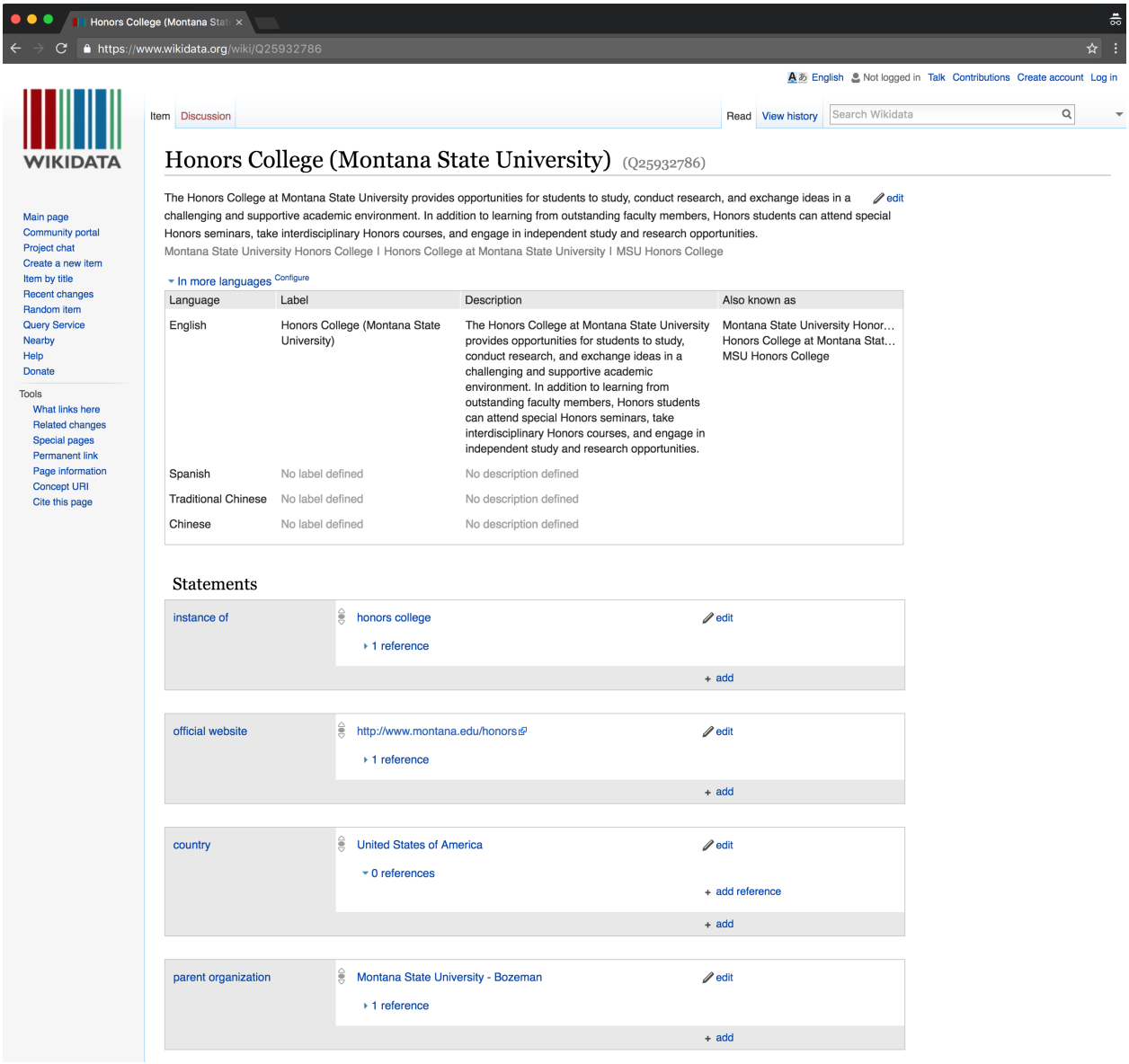

Figure 41: Portion of Wikidata record for MSU Honors College as of November 2016

As of November 2016, a robust KC appears for the MSU Honors College in Google SERP. The KC includes a description field with text drawn from the Wikipedia article (see Figure 42)

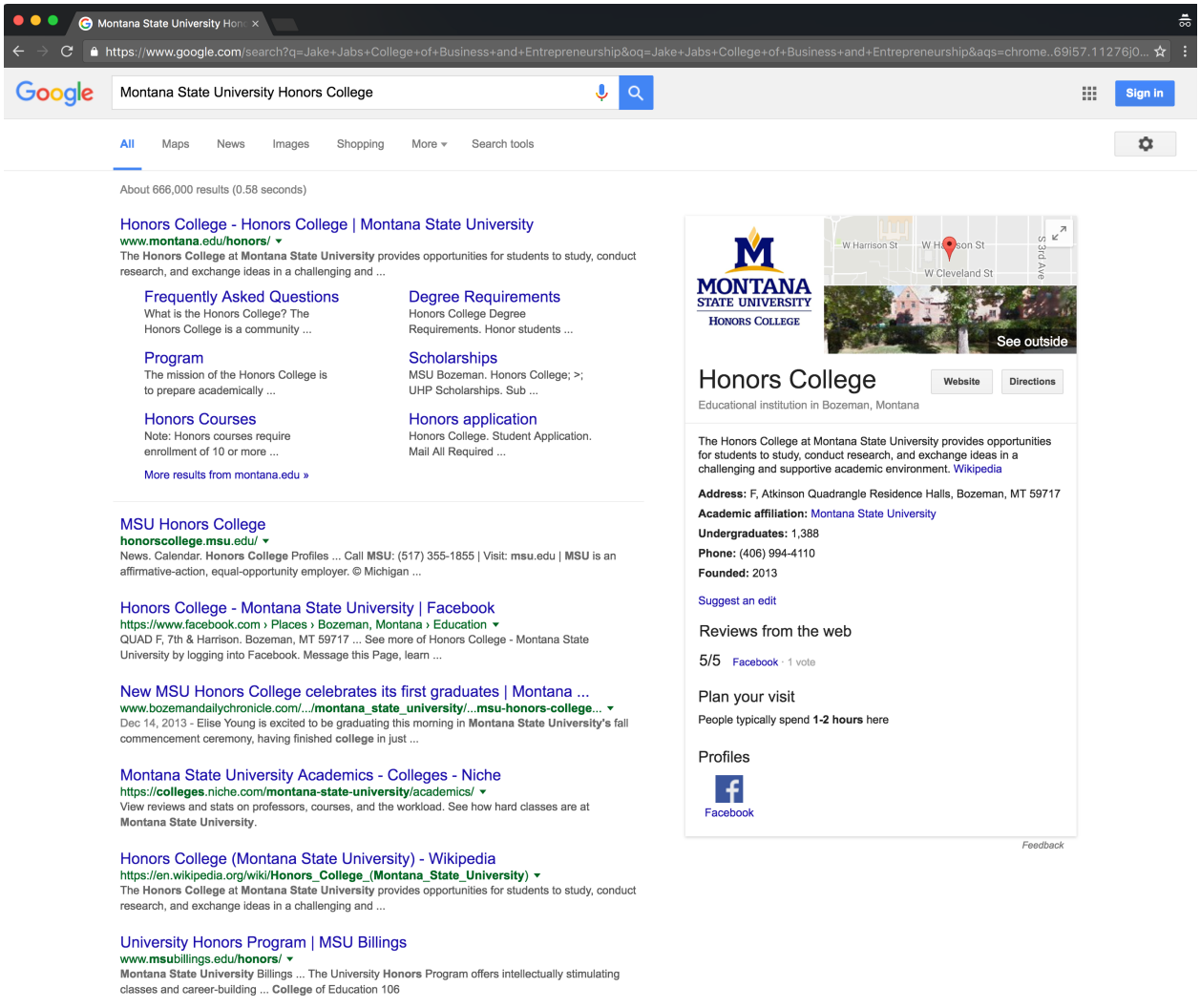

Figure 42: Google SERP in November 2016 shows KC, including description drawn from Wikipedia article

A verified Google+ profile also now exists for the MSU Honors College (see Figure 43).

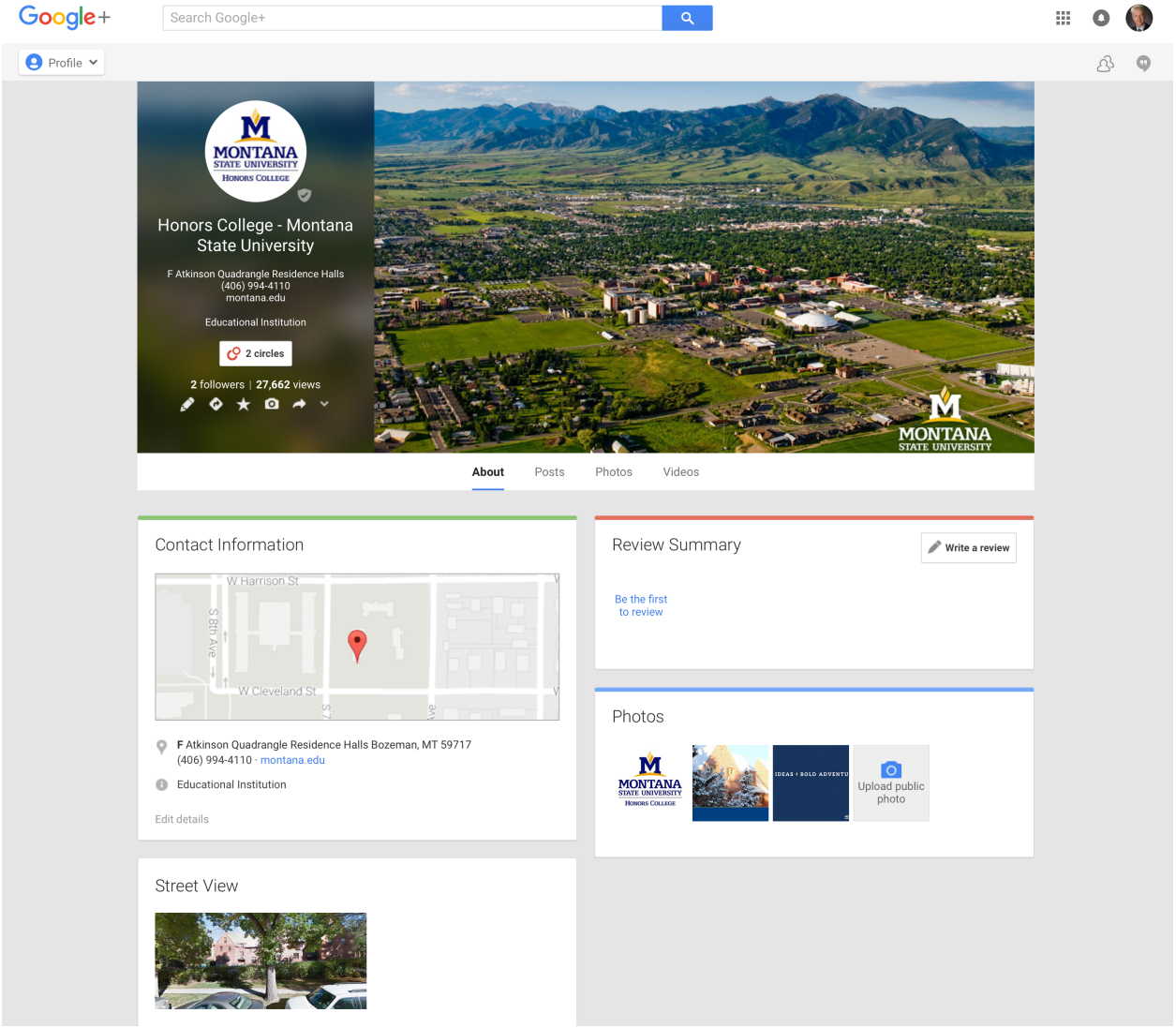

Figure 43: Verified Google+ profile for MSU Honors College in November 2016

The screen shots above show the results of the SWI service provided to the MSU Honors College by the MSU Library. None of these results could have been found prior to the SWI service that was rendered. As of this writing in November 2016 a DBpedia record still does not exist for the Honors College, but this fits with the annual cycle of DBpedia data dumps, which typically occur in the fall; the resulting linked data set is published the following spring. Since the Wikipedia article was only published in July 2016, a record should

appear in DBpedia in spring 2016, when the newest linked data set is published on DBpedia's website[6].

Section 6.5 Summary

This chapter has provided context for the broader implications of the research conducted for this dissertation. While the main data set for this study focused on a discrete group of academic libraries, some additional evidence has been provided in this chapter to show that the problem of poor SWI extends to other academic organizations within research universities, and that the SWI service tested in the case studies can be applied to other academic organizations. Montana State University represents a typical mid-sized research university in the United States, and the SWI of its eleven colleges were examined in a limited data-gathering effort. Screen capture files show the presence or lack of KC in Google SERP, and the five knowledge bases were also searched for records, as before. Results confirm that most of these organizations also lacked KC, some of the KC that did display showed inaccurate information, and only a few records for the colleges were evident in the knowledge bases.

The Action Research methodology utilized for this dissertation emphasizes using research to "solve practical problems" (Elden and Chisholm 1993). This chapter has described a new service developed by the MSU Library that aims to establish or improve the SWI of other academic organizations on the MSU campus. In its first year, the service has already demonstrated results for the first group of colleges that chose to participate; examples of the effect on the MSU Honors College were shown in this chapter. By developing the SWI service, the MSU Library has implemented the results of scholarly research to solve practical problems that improve the situation for academic organizations. Hopefully this work will encourage other academic libraries to implement similar services on their campuses.

[6] DBpedia data sets - http://wiki.dbpedia.org/datasets

Chapter 7 Conclusion

Google draws information from proprietary and LOD knowledge bases to populate its Knowledge Graph, and in turn generates Knowledge Graph Cards (KC) for entities in search results. In this study, the appearance of a KC for an academic organization has been considered an indicator of Semantic Web Identity (SWI), meaning that the search engine has assembled enough verified facts about the organization to establish it as a known entity. The data gathered and analyzed for this research support that hypothesis, but the five knowledge bases that were anticipated to influence the creation and robustness of KC seem to represent only part of the picture. Google most likely draws information from additional knowledge bases that were not included in this study, and the literature shows that it also seeks structured metadata from websites in the form of Schema.org markup. Google's trust of entities also increases if it can confirm facts about the entities in multiple sources.

Perhaps the biggest lesson that can be drawn from this research is that academic libraries lack coordinated marketing/branding strategies that are appropriate for the Semantic Web. Most ARL libraries (94/125) have more than one name, and they use those names inconsistently to communicate about their organizations. While each member library has submitted its primary name to be listed in the ARL membership directory (Baughman 2016), more go on to use their alternate names in other venues on the Web. This creates confusion in the machine-based world of the Semantic Web as search engines struggle to comprehend the name variations their crawlers find. The situation is exacerbated because most libraries have also not made any effort to establish an explicit "same as" relationship for those names in knowledge bases that afford the capability, like Google My Business (GMB) or Wikidata. Taking such action would help the machines understand that both names refer to the same organization.

The other significant lesson of this research is that ARL member libraries have been slow to engage in the knowledge bases that provide information to search engine knowledge graphs as well as to other machine consumers of linked data. The findings provide evidence that claiming and verifying a business in GMB can help generate and populate a KC, but that process has been completed for only 22% of the primary and 43% of the alternate names of the libraries. Verified Google+ profiles are automatically generated by successfully claiming and verifying a business with GMB, but at least as many Google+

profiles for the libraries have been created independently and are unverified, resulting in another marketing problem. Units and individuals within library organizations have apparently created these Google+ profiles for their library organizations, and names and facts for the organizations that appear on many of these profiles are inconsistent, even though they represent the same organization. These independently-created profiles may send confusing signals to Google. Publishing an article in Wikipedia usually results in a description appearing on the KC, but articles were evident for only 32% of library primary names and 45% of alternate names. The other three knowledge bases that were tested seem to have little or no direct influence on KC, although that may change as Wikidata has recently inherited the Freebase records that once served as a primary source for Google's Knowledge Graph. This research could not demonstrate any direct influence that DBpedia might have on SWI, but it is important to recognize that DBpedia is a significant source of linked data for other knowledge bases and services on the Semantic Web. When an organization publishes a Wikipedia article with a well-populated infobox, it offers authoritative information to direct human consumers of the text. However, the effort of researching, writing, and publishing the Wikipedia article is further leveraged when DBpedia automatically generates a rich data record from the article and makes it available to machines that consume linked data. An organization that understands this will help position itself for future syndication of DBpedia's structured data records. From simple links in websites that point to DBpedia records to help search engines understand context, to automatic generation of controlled vocabularies at the BBC (Raimond et al. 2010), and to geospatially aware applications (Becker and Bizer 2009), DBpedia is a source whose potential is just becoming known. Rather than re-keying metadata, catalog systems, CMS, DAMs, and other database could tap DBpedia for rich data records that are already linked to other records.

Why have librarians been so slow to move into the Semantic Web and why have they been so seemingly unaware of issues like SWI? A Google search for "NBA teams" (National Basketball Association), "U.S. universities," or "Rolling Stones albums" shows a Carousel display of the instances in each of those groups across the top of the screen, i.e. teams, universities, and album covers, respectively. But search for "research universities," "institutional repositories," or even "Association of Research Libraries" and no such Carousel display appears. Why? The simple technical answer is that no structured data

exists that would help Google understand that these entities are composed of instances. Even now, the DBpedia record for "Library" displays subject headings of "Book promotion; Library; and Library science," and the RDF types include terms like "Artifact; Object; Physical entity; and Structure" (see Figure 73). Is this the best machine-readable description that librarians can give of a library?

Beyond the technical reasons that Google can't adequately represent libraries and library-related issues in its SERP, there are cultural biases that have limited librarians from engaging more fully with knowledge bases and other Semantic Web data sources. Despite being the most commonly used search engine in North America and Europe, librarians have historically harbored a suspicion of Google as an information source, even actively discouraging others from using it. Some of this suspicion may have been rooted in insecurity, as librarians watched their reference users diminish (Kennedy 2011), or in a sense that Google could not possibly be scholarly enough for an academic setting. Wikipedia was also the target of librarian derision, although that stance seems to have softened in recent years as the Wikimedia Foundation and groups like OCLC, Inc. have promoted its use through programs like the Wikipedia Visiting Scholars program (Stinson and Orlowitz 2015). But tolerance is a long way from proactive engagement, and many librarians don't seem to understand that Wikipedia is not just an encyclopedia for human-readable text. Beyond Google and Wikipedia, it is probable that most librarians simply don't know enough about the Semantic Web and its potential to put machines to work breaking down the siloes of informational wealth in academic disciplines, and connecting seemingly disconnected data points that could lead to advances in all manner of knowledge. Whatever the cause, the time is late for librarians to proactively engage with Semantic Web data sources that are helping machines to connect humans to information.

While studying the methods to make a KC appear for an organization is interesting technical work, that exercise alone misses the larger point. In a time when academic libraries feel pressure to articulate their value proposition, this research shows that they could provide a service that creates and maintains SWI for organizations across their institutions, possibly with considerable effect. Academic libraries can develop and offer SWI services for their campus organizations by adapting traditional skills in cataloging and metadata, as well as in research and scholarly publishing.

Learning to establish and maintain the SWI of their own organizations is prerequisite for libraries to help position them as experts on the topic. SWI for academic organizations hinges, above all, on a clear branding strategy by the organization and on technical efforts to minimize confusion for the machines that read data on the Semantic Web. Neglecting SWI, or engaging only haphazardly in sources that help create SWI can result in a presence on the Web that is detrimental to the representation of the organization. The research in this dissertation has shown cases where enterprising individuals in organizations have taken it upon themselves to create Wikipedia articles, Google+ profiles, Wikidata records, or even GMB profiles, but without direction from the organization's administrators and therefore without a clear understanding of how the organization wants to be represented. This has sometimes resulted in multiple profiles showing different addresses, hours, logos, and even different names. The phenomenon demonstrates a disconnect between library administrators, who may not understand Semantic Web technologies and data sources very well, and the librarians or library staff who have a better understanding and are not waiting for their administrators to catch up. Academic library administrators must understand the basic process of establishing SWI and its potential effect so that they can drive SWI as a strategic initiative for their own organizations as well as campus constituents. This is not a difficult subject; it's just that awareness is lacking.

There is much more research that could be conducted with the data set that was compiled for this research, or with an updated or expanded version of the data. Other knowledge bases could be added to the study, such as the CIA World Factbook and OpenCyc, and an effort could be made to monitor SWI results from Schema.org markup that is embedded in the HTML pages on organizations' websites. While this study has focused on Google and its KC, additional testing could be conducted with Bing, as it also generates a KC when it has established enough facts concerning an entity. Much more data about the effects of SWI can be collected through the "insights" offered by GMB and through other analytics tools.

Universities spend a great deal of money on marketing staff and campaigns, and the question could rightfully be asked whether university communications offices are the more appropriate curators of SWI for their institutions. Currently, there is a vacuum in the responsibility for SWI because it is not very well understood and few, if any, academic organizations can formally address the problem. In the continued absence of an alternative

there is little doubt that communications offices will soon begin to take on the role of establishing and maintaining SWI. This development, if it were to occur, would represent a missed opportunity for academic libraries, and probably would also fail to bring the appropriate skills to the problem. Librarians have long been involved with creating and curating structured data records, and SWI as described in this study represents exactly that. Non-librarians would have a harder time understanding structured data, including the numerous controlled vocabularies and metadata schema that are available.

Establishing SWI is not an exact science and probably never will be. There is no certain formula at this point that will result in accurate and robust KC to better represent academic organizations and that would indicate a search engine understands the existence and intent of those organizations. The processes described in this research should be considered as indicative rather than prescriptive. Trying to pinpoint which knowledge bases should be populated is much less important than an overall awareness of the growing importance of LOD and other knowledge bases from which search engines may draw to build their knowledge graphs. While the research in this dissertation focused on Google and its related products, the concepts should be adaptable to other semantic search engines.

References

Aghaei, Sareh, Mohammad Ali Nematbakhsh, and Hadi Khosravi Farsani. 2012. "Evolution of the World Wide Web: From Web 1.0 to Web 4.0." *International Journal of Web & Semantic Technology* 3 (1): 1–10.

Alexa Internet, Inc. 2016. "Wikipedia.org Traffic Statistics." Commercial. *Alexa*. http://www.alexa.com/siteinfo/wikipedia.org.

Ali, Mohamed, and T. Padma. 2016. "Graph Database: A Contemporary Storage Mechanism for Connected Data." *International Journal of Advanced Research in Computer and Communication Engineering* 5 (3): 930–33.

Alkindi, Salim Said, and Mohammed Nasser Al-Suqri. 2013. "Social Networking Sites as Marketing and Outreach Tools of Library and Information Services." *Global Journal of Human Social Science, Arts, Humanities & Psychology* 13 (2): 15.

Alphabet, Inc. 2015. "Consolidated Revenues." Form 10K. Washington, D.C.: United States Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1288776/000165204416000012/goog10-k2015.htm#s2A481E6E5C511C2C8AAECA5160BB1908.

Anderson, Rick. 2005. "The (Uncertain) Future of Libraries in a Google World: Sounding an Alarm." *Internet Reference Services Quarterly* 10 (3–4): 29–36. http://dx.doi.org/10.1300/J136v10n03_04.

Angles, Renzo, and Claudio Gutierrez. 2008. "Survey of Graph Database Models." *ACM Computing Surveys* 40 (1): 1–39.

ARL Board. 2013. "Principles of Membership in the Association of Research Libraries." Association of Research Libraries. http://www.arl.org/storage/documents/publications/membership-principles-2013revision-final.pdf.

Arlitsch, Kenning. 2014a. "Being Irrelevant: How Library Data Interchange Standards Have Kept Us Off the Internet." *Journal of Library Administration* 54 (7): 609–19. http://dx.doi.org/10.1080/01930826.2014.964031.

———. 2014b. "Exposing Library Collections on the Web: Challenges and Lessons Learned." Videorecording presented at the CNI: Coalition for Networked Information Fall 2014

Membership Meeting, Washington, D.C., December 8. https://www.youtube.com/watch?v=WEl0CJPI4DI.

———. 2015. “Walk Before You Run: Prerequisites to Linked Data.” PowerPoint presented at the Linked Data & RDF: New Frontiers in Metadata and Access, Amigos Library Services online conference, April 23. http://www.slideshare.net/karlitsch/walk-before-you-run-prerequisites-to-linked-data.

———. 2016. “Data Set Supporting the Dissertation ‘Semantic Web Identity in Academic Organizations: Search Engine Entity Recognition and the Sources That Influence Knowledge Graph Cards in Search Results.’” Montana State University ScholarWorks. https://doi.org/10.15788/M2F590.

Arlitsch, Kenning, Patrick OBrien, Jason A. Clark, Doralyn Rossmann, Scott W. H. Young, and Leila Sterman. 2015. “Emphasizing Institutional Identity: Applying Semantic Web Optimization to Improve MSU’s Presence on the Web.” https://drive.google.com/open?id=0B6oeJCZ1Vix6SDVHOTNySDIwWEU.

Arlitsch, Kenning, Patrick OBrien, Jason A. Clark, Scott W. H. Young, and Doralyn Rossmann. 2014. “Demonstrating Library Value at Network Scale: Leveraging the Semantic Web with New Knowledge Work.” *Journal of Library Administration* 54 (5): 413–25. http://dx.doi.org/10.1080/01930826.2014.946778.

Arlitsch, Kenning, Patrick OBrien, and Brian Rossmann. 2013. “Managing Search Engine Optimization: An Introduction for Library Administrators.” *Journal of Library Administration* 53 (2–3): 177–88. http://dx.doi.org/10.1080/01930826.2013.853499.

Arlitsch, Kenning, and Patrick S. O’Brien. 2012. “Invisible Institutional Repositories: Addressing the Low Indexing Ratios of IRs in Google Scholar.” *Library Hi Tech* 30 (1): 60–81. http://dx.doi.org/10.1108/07378831211213210.

Arlitsch, Kenning, and Patrick S OBrien. 2013. *Improving the Visibility and Use of Digital Repositories through SEO*. Chicago: ALA TechSource, an imprint of the American Library Association. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=578551.

Assaf, Ahmad, Ghislain A. Atemezing, Raphael Troncy, and Elena Cabrio. 2014. “What Are the Important Properties of an Entity? Comparing Users and Knowledge Graph Point of View.” In *The Semantic Web: ESWC 2014 Satellite Events*, 8798:190–94. Cham:

Springer International Publishing. http://link.springer.com/10.1007/978-3-319-11955-7_76.

Association of Research Libraries. 2016. “List of ARL Members.” Non-profit. *Association of Research Libraries*. http://www.arl.org/membership/list-of-arl-members.

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. “DBpedia: A Nucleus for a Web of Open Data.” In *The Semantic Web*, 4825:722–35. Berlin, Heidelberg: Springer Berlin Heidelberg. http://link.springer.com/10.1007/978-3-540-76298-0_52.

Baskerville, Richard L. 1999. “Investigating Information Systems with Action Research.” *Communications of the Association for Information Systems* 2 (3es). http://www.uio.no/studier/emner/matnat/ifi/INF9930/v12/undervisningsmateriale/Baskerville-1999-IS-Action-Research.pdf.

Baskerville, Richard L., and A.T. Wood-Harper. 1996. “A Critical Perspective on Action Research as a Method for Information Systems Research.” *Journal of Information Technology* 11 (3): 235–46.

Baughman, Sue. 2016. “(Email) Re: Question about ARL Member Listings,” July 27.

Becker, Christian, and Christian Bizer. 2009. “Exploring the Geospatial Semantic Web with DBpedia Mobile.” *Web Semantics: Science, Services and Agents on the World Wide Web* 7 (4): 278–86. http://dx.doi.org/10.1016/j.websem.2009.09.004.

Bergman, Mike. 2012. “Deconstructing the Google Knowledge Graph.” *AI3: Adaptive Information, Adaptive Innovation, Adaptive Infrastructure*. May 18. http://www.mkbergman.com/1009/deconstructing-the-google-knowledge-graph/.

Berners-Lee, Tim. 1996. “WWW: Past, Present, and Future.” *Computer* 29 (10): 69–77. http://dx.doi.org/10.1109/2.539724.

———. 2006. “Linked Data.” *W3C*. July 27. https://www.w3.org/DesignIssues/LinkedData.html.

Berners-Lee, Tim, Robert Cailliau, Ari Luotonen, Henrik Frystyk Nielsen, and Arthur Secret. 1994. “The World-Wide Web.” *Communications of the ACM* 37 (8): 76–82. http://dx.doi.org/10.1145/179606.179671.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. “The Semantic Web.” *Scientific American* 284 (5): 34–43. http://dx.doi.org/10.1038/scientificamerican0501-34.

Bernstein, Michael S., Jaime Teevan, Susan Dumais, Daniel Liebling, and Eric Horvitz. 2012. "Direct Answers for Search Queries in the Long Tail." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 237. Austin, TX: ACM Press. http://dx.doi.org/10.1145/2207676.2207710.

Bizer, Christian, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. "DBpedia - A Crystallization Point for the Web of Data." *Web Semantics: Science, Services and Agents on the World Wide Web* 7 (3): 154–65. http://dx.doi.org/10.1016/j.websem.2009.07.002.

Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. "Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge." In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, 1247. ACM Press. http://dx.doi.org/10.1145/1376616.1376746.

Bouquet, Paolo, Heiko Stoermer, and Massimiliano Vignolo. 2012. "Web of Data and Web of Entities: Identity and Reference in Interlinked Data in the Semantic Web." *Philosophy & Technology* 25 (1): 5–26. http://dx.doi.org/10.1007/s13347-010-0011-6.

Brin, Sergey, and Lawrence Page. 1998. "The Anatomy of a Large-Scale Hypertextual Web Search Engine." *Computer Networks and ISDN Systems* 30 (1–7): 107–17. http://dx.doi.org/10.1016/S0169-7552(98)00110-X.

Bruehmmer, Paul. 2013. "Future SEO: String Entity Optimization." *Search Engine Land*. September 9. http://searchengineland.com/killer-seo-string-entity-optimization-171094?utm_source=gplus&utm_medium=social&utm_campaign=pluspost.

Brutlag, Jake. 2009. "Speed Matters." *Google Research Blog*. June 23. http://googleresearch.blogspot.com/2009/06/speed-matters.html.

Bryant, Martin. 2011. "20 Years Ago Today, the World Wide Web Opened to the Public." Commercial. *TNW Insider*. August 6. http://thenextweb.com/insider/2011/08/06/20-years-ago-today-the-world-wide-web-opened-to-the-public/.

Bureau of Labor Statistics. 2016. "Other Information Services: NAICS 519." Government. *United States Department of Labor*. November 18. http://www.bls.gov/iag/tgs/iag519.htm.

Butzbach, Alex. 2014. “Freebase Is Shutting down - What Does It Mean for the Knowledge Graph and SEO?” *Brafton: Fuel Your Brand*. December 9. http://www.brafton.com/news/freebase-shutting-mean-knowledge-graph-seo/.

Cahill, Kay, and Renee Chalut. 2009. “Optimal Results: What Libraries Need to Know About Google and Search Engine Optimization.” *The Reference Librarian* 50 (3): 234–47. http://dx.doi.org/10.1080/02763870902961969.

Callahan, Ewa S., and Susan C. Herring. 2011. “Cultural Bias in Wikipedia Content on Famous Persons.” *Journal of the American Society for Information Science and Technology* 62 (10): 1899–1915. http://dx.doi.org/10.1002/asi.21577.

Cals, Jochen WL, and Daniel Kotz. 2008. “Researcher Identification: The Right Needle in the Haystack.” *The Lancet* 371 (9631): 2152–53. http://dx.doi.org/10.1016/S0140-6736(08)60931-9.

Carnduff, Brett. 2014. “‘Google My Business’ vs. ‘Google+ Page’ - What’s the Difference?” *Brent Carnduff --- for the Accidental SEO*. December 22. http://brentcarnduff.com/google-business-vs-google-page-whats-difference/.

Carter, Troy M., and Priscilla Seaman. 2011. “The Management and Support of Outreach in Academic Libraries.” *Reference & User Services Quarterly* 51 (2): 163–71.

“Category: Infobox Templates.” 2016. Non-profit. *Wikipedia, the Free Encyclopedia*. March 29. https://en.wikipedia.org/wiki/Category:Infobox_templates.

Center for Postsecondary Research. 2016. “2015 Update, Facts & Figures: Descriptive Highlights.” The Carnegie Classification of Institutions of Higher Education. http://carnegieclassifications.iu.edu/downloads/CCIHE2015-FactsFigures.pdf.

Central Intelligence Agency. 2015. “CIA World Factbook.” Government. *The World Factbook*. https://www.cia.gov/library/publications/the-world-factbook/.

Chernatony, Leslie de. 2002. “Living the Corporate Brand: Brand Values and Brand Enactment.” *Corporate Reputation Review* 5 (2/3): 113–32.

Christensen, Clayton M. 1997. *The Innovator’s Dilemma: When New Technologies Cause Great Firms to Fail*. The Management of Innovation and Change Series. Boston, Mass: Harvard Business School Press.

comScore, Inc. 2016. “comScore Releases February 2016 U.S. Desktop Search Engine Rankings.” Commercial. *comScore*. March 16.

https://www.comscore.com/Insights/Rankings/comScore-Releases-February-2016-US-Desktop-Search-Engine-Rankings.

Connaway, Lynn Silipigni, and Timothy J. Dickey. 2010. "The Digital Information Seeker: Report of Findings from Selected OCLC, RIN, and JISC User Behavior Projects." Dublin, Ohio: OCLC Research. http://www.jisc.ac.uk/media/documents/publications/reports/2010/digitalinformationseekerreport.pdf.

Connaway, Lynn Silipigni, and Ronald R. Powell. 2010. *Basic Research Methods for Librarians*. 5th ed. Santa Barbara, CA: Libraries Unlimited. https://books.google.com/books?hl=en&lr=&id=_7ySMR0T9uYC&oi=fnd&pg=PP1&dq=Basic+research+methods+for+librarians&ots=5JAHmYq4oT&sig=0MQRg4jU8iaEP7d1RVFn899QPMg#v=onepage&q=Basic%20research%20methods%20for%20librarians&f=false.

Cook, Douglas. 2011. "The Recursive Cycle of Qualitative Action Research." In *Using Qualitative Methods in Action Research: How Librarians Can Get to the Why of Data*, 252 pp. Chicago: Association of College and Research Libraries.

Crosetto, Alice, and Thomas A Atwood. 2012. "Naming Academic Libraries: Is Institutional Identity Obscuring the Generous Benefactors and Illustrious Educators of Old?" *Names* 60 (2): 90–104. http://dx.doi.org/10.1179/0027773812Z.00000000012.

Cunningham, J.Barton. 1993. *Action Research and Organizational Development*. Westport, CT: Praeger.

Dalgaard, Peter. 2002. *Introductory Statistics with R*. Statistics and Computing. New York: Springer.

Dame, Nate. 2015. "What Can Businesses Do about the Knowledge Graph Dominating Search Results?" Commercial. *Search Engine Land*. February 15. http://searchengineland.com/businesses-knowledge-graph-dominating-search-results-215520.

Davis, Gordon B. 2006. "Information Systems as an Academic Discipline." In *The Past and Future of Information Systems: 1976–2006 and Beyond*, edited by David Avison, Steve Elliot, John Krogstie, and Jan Pries-Heje, 214:11–25. Springer US. http://link.springer.com/10.1007/978-0-387-34732-5_2.

Davison, Robert, Maris G. Martinsons, and Ned Kock. 2004. “Principles of Canonical Action Research.” *Information Systems Journal* 14 (1): 65–86. http://dx.doi.org/10.1111/j.1365-2575.2004.00162.x.

“DBpedia.” 2013. *Freebase*. Google, Inc.

Dempsey, Lorcan. 2014. *The Network Reshapes the Library: Lorcan Dempsey on Libraries, Services and Networks*. Edited by Kenneth J. Varnum. Chicago: ALA Editions, an imprint of the American Library Association.

Dennis, Melissa. 2012. “Outreach Initiatives in Academic Libraries, 2009-2011.” *Reference Services Review* 40 (3): 368–83. http://dx.doi.org/10.1108/00907321211254643.

DePianto, Susie. 2016. “Helping Prospective Students Make Decisions about Their Future.” *Google Education Research*. September 30. https://blog.google/topics/education/helping-prospective-students-make-decisions-about-their-future/.

DeRidder, Jody L. 2008. “Googlizing a Digital Library.” *The Code4Lib Journal*, no. 2 (March). http://journal.code4lib.org/articles/43.

Dickens, L., and K. Watkins. 1999. “Action Research: Rethinking Lewin.” *Management Learning* 30 (2): 127–40. http://dx.doi.org/10.1177/1350507699302002.

Dou, Wenyu, Kai H. Lim, Chenting Su, Nan Zhou, and Nan Cui. 2010. “Brand Positioning Strategy Using Search Engine Marketing.” *MIS Quarterly* 34 (2): 261-A4.

Drakos, Nikos, Marcus Hennecke, Ross Moore, and Herb Swan. 2005. “Logistic Regression.” Educational. December 21. http://pages.uoregon.edu/aarong/teaching/G4075_Outline/node16.html.

Duignan, Brian. 2015. “Empiricism.” *Encyclopaedia Brittanica*. Encyclopaedia Britannica, Inc. http://www.britannica.com/topic/empiricism.

Easterbrook, Steve, Janice Singer, Margaret-Anne Storey, and Daniela Damian. 2008. “Selecting Empirical Methods for Software Engineering Research.” In *Guide to Advanced Empirical Software Engineering*, edited by Forrest Shull, Janice Singer, and Dag I. K. Sjøberg, 285–311. London: Springer London. http://link.springer.com/10.1007/978-1-84800-044-5_11.

Edward, Tony. 2015. “Leveraging Wikidata to Gain a Google Knowledge Graph Result.” *Search Engine Land*. May 1. http://searchengineland.com/leveraging-wikidata-gain-google-knowledge-graph-result-219706.

———. 2016. “How to Enhance Your Google Knowledge Graph Result (Case Study).” *Search Engine Land*. April 1. http://searchengineland.com/enhance-google-knowledge-graph-result-case-study-243965.

Elden, M., and Rupert F. Chisholm. 1993. “Emerging Varieties of Action Research: Introduction to the Special Issue.” *Human Relations* 46 (2): 121–42. http://dx.doi.org/10.1177/001872679304600201.

Elson Anderson, Katie, and Julie M. Still. 2011. “An Introduction to Google Plus.” *Library Hi Tech News* 28 (8): 7–10. http://dx.doi.org/10.1108/07419051111187842.

Enserink, M. 2009. “SCIENTIFIC PUBLISHING: Are You Ready to Become a Number?” *Science* 323 (5922): 1662–64. http://dx.doi.org/10.1126/science.323.5922.1662.

Ert, Eyal, Aliza Fleischer, and Nathan Magen. 2016. “Trust and Reputation in the Sharing Economy: The Role of Personal Photos in Airbnb.” *Tourism Management* 55 (August): 62–73. http://dx.doi.org/10.1016/j.tourman.2016.01.013.

Erxleben, Fredo, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. “Introducing Wikidata to the Linked Data Web.” In *The Semantic Web – ISWC 2014*, edited by Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble, 8796:50–65. Cham: Springer International Publishing. http://link.springer.com/10.1007/978-3-319-11964-9_4.

Fabian, Carole Ann, Charles D’aniello, Cynthia Tysick, and Michael Morin. 2003. “Multiple Models for Library Outreach Initiatives.” *The Reference Librarian* 39 (82): 39–55. http://dx.doi.org/10.1300/J120v39n82_04.

Farmer, Lesley. 2011. “Qualitative Research and the Librarian.” In *Using Qualitative Methods in Action Research: How Librarians Can Get to the Why of Data*, 252 pp. Chicago: Association of College and Research Libraries.

Farrar, Donald E., and Robert R. Glauber. 1967. “Multicollinearity in Regression Analysis: The Problem Revisited.” *The Review of Economics and Statistics* 49 (1): 92. http://dx.doi.org/10.2307/1937887.

Freudenberg, Markus, Dimitris Kontokostas, and Sebastian Hellmann. 2016. “DBpedia Version 2015-10.” *DBpedia*. http://wiki.dbpedia.org/dbpedia-dataset-version-2015-10.

Galliers, Robert D., and Frank F. Land. 1987. "Viewpoint: Choosing Appropriate Information Systems Research Methodologies." *Communications of the ACM* 30 (11): 901–2. http://dx.doi.org/10.1145/32206.315753.

Gentleman, Robert, and Ross Ihaka. 2016. *R* (version 3.2.4). X86_64-apple-darwin13.4.0 (64-bit). En. R Foundation. https://www.r-project.org.

George, Lee Anne, and Julia C. Blixrud. 2002. *Celebrating Seventy Years of the Association of Research Libraries, 1932-2002*. 2008th ed. Washington, DC: Association of Research Libraries. http://www.arl.org/storage/documents/publications/celebrating-seventy-years-arl.pdf.

Gesenhues, Amy. 2013. "Google Officially Launches Knowledge Graph Carousel for Local Search." Commercial. *Search Engine Land*. June 18. http://searchengineland.com/google-officially-launches-knowledge-graph-carousel-for-local-search-163809.

Gill, Bradley C., Anna M. Zampini, and Neil B. Mehta. 2015. "Digital Identity: Develop One Before You're Given One." *Urology* 85 (6): 1219–23. http://dx.doi.org/10.1016/j.urology.2015.02.056.

Google, Inc. 2016a. "Browse in Private with Incognito Mode." Commercial. *Chrome Help*. https://support.google.com/chrome/answer/95464?hl=en.

———. 2016b. "Get Your Ad on Google Today." Commercial. *Google AdWords*. https://adwords.google.com.

———. 2016c. "Request a Change to a Knowledge Graph Card in Search Results." *Google Search Help*. https://support.google.com/websearch/answer/6325583?p=kg_edit&rd=1.

Google Knowledge Graph Team. 2014. "Freebase: News and Tips on Using Freebase." Commercial. *Google+*. December 16. https://plus.google.com/109936836907132434202/posts/bu3z2wVqcQc?cfem=1.

Guttentag, Daniel. 2015. "Airbnb: Disruptive Innovation and the Rise of an Informal Tourism Accommodation Sector." *Current Issues in Tourism* 18 (12): 1192–1217. http://dx.doi.org/10.1080/13683500.2013.827159.

Haak, Laurel L., Martin Fenner, Laura Paglione, Ed Pentz, and Howard Ratner. 2012. "ORCID: A System to Uniquely Identify Researchers." *Learned Publishing* 25 (4): 259–64. http://dx.doi.org/10.1087/20120404.

Hagans, Andy. 2005. “High Accessibility Is Effective Search Engine Optimization.” *A List Apart*, November 8. http://www.alistapart.com/articles/accessibilityseo.

Hamari, Juho, Mimmi Sjöklint, and Antti Ukkonen. 2015. “The Sharing Economy: Why People Participate in Collaborative Consumption.” *Journal of the Association for Information Science and Technology*, July, n/a-n/a. http://dx.doi.org/10.1002/asi.23552.

Heaton, James. 2011. “The Difference between Marketing and Branding.” Commercial. *Tronvig Group*. http://www.tronviggroup.com/the-difference-between-marketing-and-branding/.

Hepburn, P., and K. M. Lewis. 2008. “What’s in a Name? Using Card Sorting to Evaluate Branding in an Academic Library’s Web Site.” *College & Research Libraries* 69 (3): 242–51. http://dx.doi.org/10.5860/crl.69.3.242.

Higginbottom, Patricia C., and Valerie S. Gordon. 2016. *Marketing for Special and Academic Libraries: A Planning and Best Practices Sourcebook*. Medical Library Association Books. Lanham: Rowman & Littlefield.

Hosmer, David W., and Stanley Lemeshow. 2000. *Applied Logistic Regression*. 2nd ed. Wiley Series in Probability and Statistics. New York: Wiley.

Izenstark, Amanda. 2014. “Look Good When You’re Googled: Creating and Optimizing Your Digital Identity.” *Library Hi Tech News* 31 (9): 14–16. http://dx.doi.org/10.1108/LHTN-07-2014-0061.

Jackson, Todd. 2010. “Introducing Google Buzz.” Commercial. *Google Official Blog*. February 9. https://googleblog.blogspot.com/2010/02/introducing-google-buzz.html.

Jentzsch, Anja. 2009. “DBpedia - Extracting Structured Data from Wikipdia.” PDF presented at the Semantic Web in Bibliotheken, Cologne, Germany. http://www.anjajentzsch.de/slides/SWIB09_DBpedia.pdf.

Jing (version 2.7.0). 2014. En. Okemos, MI: TechSmith Corporation. http://www.techsmith.com/jing.html.

Johnson, L, S Becker, V Estrada, and A Freeman. 2014. “The NMC Horizon Report: 2014 Higher Education Edition.” ISBN 978-0-9897335-5-7. New Media Consortium. http://www.nmc.org/pdf/2014-nmc-horizon-report-he-EN.pdf.

Kennedy, Scott. 2011. “Farewell to the Reference Librarian.” *Journal of Library Administration* 51 (4): 319–25. http://dx.doi.org/10.1080/01930826.2011.556954.

Khabsa, Madian, and C. Lee Giles. 2014. “The Number of Scholarly Documents on the Public Web.” Edited by Ren Zhang. *PLoS ONE* 9 (5): e93949. http://dx.doi.org/10.1371/journal.pone.0093949.

Kohli, Chiranjeev, and Douglas W. LaBahn. 1995. “Creating Effective Brand Names: A Study of the Naming Process.” ISBM REport 12-1995. Institute for the Study of Business Markets. University Park, PA: Pennsylvania State University. https://web.wpi.edu/Pubs/E-project/Available/E-project-121510-165023/unrestricted/Creating_an_Effective_Brand_Name.pdf.

Kunder, Maurice de. 2016. “The Size of the World Wide Web (The Internet).” *WorldWideWebSize.com*. October 2. http://www.worldwidewebsize.com.

Lalithsena, Sarasi, Pascal Hitzler, Amit Sheth, and Prateek Jain. 2013. “Automatic Domain Identification for Linked Open Data.” In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 205–12. Atlanta, GA: IEEE. http://dx.doi.org/10.1109/WI-IAT.2013.206.

Lawler, Ryan. 2013. “In an Effort to Connect Users’ Online and Offline Identities, Airbnb Introduces Verified Identification.” *TechCrunch*. April 30. https://techcrunch.com/2013/04/30/airbnb-verified-id/.

Lehmann, Jens, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, et al. 2015. “DBpedia - A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia.” *Semantic Web* 6 (2): 167–195. http://dx.doi.org/10.3233/SW-140134.

Lewin, Kurt. 1946. “Action Research and Minority Problems.” *Journal of Social Issues* 2 (4): 34–46. http://dx.doi.org/10.1111/j.1540-4560.1946.tb02295.x.

Li, K., Y. Li, Y. Zhou, Z. Lv, and Y. Cao. 2013. Knowledge-based entity detection and disambiguation. USPTO US20130173604 A1, filed December 30, 2011, and issued July 2013. https://www.google.com/patents/US20130173604.

Lih, Andrew. 2009. *The Wikipedia Revolution: How a Bunch of Nobodies Created the World’s Greatest Encyclopedia*. 1st ed. New York: Hyperion.

Lloyd, Dave. 2014. “What’s next for SEO? Semantic Web Optimization.” *Digital Marketing Blog*. May 7. http://blogs.adobe.com/digitalmarketing/search-marketing/whats-next-seo-semantic-web-optimization/.

Luyt, Brendan, Yasmin Ally, Nur Hakim Low, and Norah Binte Ismail. 2010. "Librarian Perception of Wikipedia: Threats or Opportunities for Librarianship?" *Libri* 60 (1): 57–64. http://dx.doi.org/10.1515/libr.2010.005.

MacColl, John. 2006. "Google Challenges for Academic Libraries." *Ariadne*, no. 46 (February). http://www.ariadne.ac.uk/issue46/maccoll/.

Manyika, James, and Charles Roxburgh. 2011. "The Great Transformer: The Impact of the Internet on Economic Growth and Prosperity." McKinsey Global Institute. http://www.mckinsey.com/industries/high-tech/our-insights/the-great-transformer.

Matthews, Brian. 2005. "Semantic Web Technologies." *E-Learning*, JISC Technology and Standards Watch, 6 (6).

Meij, Edgar, Krisztian Balog, and Daan Okijk. 2014. "Entity Linking and Retrieval for Semantic Search." PDF presented at the WSDM, New York City, February 24. https://raw.githubusercontent.com/ejmeij/entity-linking-and-retrieval-tutorial/master/MTL%202014/20140615%20Entity%20Linking%20and%20Retrieval%20for%20Semantic%20Search%20Tutorial%20-%2001%20Introduction.pdf.

Mendes, Pablo, and Max Jakob. 2012. "Who's New in Google Summer of Code 2012: Part 3." *Google Open Source Blog*. August 24. http://google-opensource.blogspot.co.uk/2012/08/whos-new-in-google-summer-of-code-2012_24.html.

Messner, Marcus, and Marcia W. DiStaso. 2013. "Wikipedia versus Encyclopedia Britannica: A Longitudinal Analysis to Identify the Impact of Social Media on the Standards of Knowledge." *Mass Communication and Society* 16 (4): 465–86. http://dx.doi.org/10.1080/15205436.2012.732649.

Meyer, Robinson. 2013. "90% of Wikipedia's Editors Are Male - Here's What They're Doing about It." *The Atlantic*, October 25.

———. 2015. "Europeans Use Google Way, Way More than Americans Do: Google's Huge Market Share Is Part of What Strengthens the EU's Antitrust Case." *The Atlantic*, April 15. http://www.theatlantic.com/technology/archive/2015/04/europeans-use-google-way-way-more-than-americans-do/390612/.

Mingers, John. 2003. "The Paucity of Multimethod Research: A Review of the Information Systems Literature." *Information Systems Journal* 13 (3): 233–49. http://dx.doi.org/10.1046/j.1365-2575.2003.00143.x.

Morrison, Alan. 2013. “What Is the Difference between Wikidata and DBpedia?” *Quora*. March 7. http://www.quora.com/What-is-the-difference-between-Wikidata-and-DBpedia.

Nickolai, Daniel H., Steve G. Hoffman, and Mary Nell Trautner. 2012. “Can a Knowledge Sanctuary Also Be an Economic Engine? The Marketization of Higher Education as Institutional Boundary Work: Marketization of Higher Education as Institutional Boundary Work.” *Sociology Compass* 6 (3): 205–18. http://dx.doi.org/10.1111/j.1751-9020.2011.00449.x.

Nielsen, Peter Axel. 2007. “IS Action Research and Its Criteria.” In *Information Systems Action Research*, edited by Ned Kock, 13:355–75. Boston, MA: Springer US. http://link.springer.com/10.1007/978-0-387-36060-7_15.

Orsini, Lauren. 2014. “The Rise and Fall of Orkut: Google’s Decade-Long Social Media Experiment.” Commercial. *ReadWrite*. June 30. http://readwrite.com/2014/06/30/the-rise-and-fall-of-orkut-googles-decade-long-social-media-experiment/.

Ovadia, Steven. 2011. “An Early Introduction to the Google+ Social Networking Project.” *Behavioral & Social Sciences Librarian* 30 (4): 259–63. http://dx.doi.org/10.1080/01639269.2011.622258.

Palmer, Adrian. 2009. *Introduction to Marketing: Theory and Practice*. 2nd ed. Oxford ; New York: Oxford University Press.

Panda, Tapan K. 2013. “Search Engine Marketing: Does the Knowledge Discovery Process Help Online Retailers?” *IUP Journal of Knowledge Management* 11 (3): 56–66.

Patel, Neil. 2015. “The Beginner’s Guide to Google’s Knowledge Graph.” Commercial. *NeilPatel*. June 30. http://neilpatel.com/2015/06/30/the-beginners-guide-to-the-googles-knowledge-graph/.

Perrott, James. 2015. “Google Answer Boxes: The What, Why and How.” Commercial. *Search Engine Watch*. June 24. https://searchenginewatch.com/sew/opinion/2414342/google-answer-boxes-the-what-why-and-how.

Pfisterer, Dennis, Kay Romer, Daniel Bimschas, Oliver Kleine, Richard Mietz, Cuong Truong, Henning Hasemann, et al. 2011. “SPITFIRE: Toward a Semantic Web of Things.” *IEEE*

Communications Magazine 49 (11): 40–48. http://dx.doi.org/10.1109/MCOM.2011.6069708.

Pirsig, Robert M. 1974. *Zen and the Art of Motorcycle Maintenance: An Inquiry into Values*. New York: Morrow.

Raimond, Yves, Tom Scott, Patrick Sinclair, Libby Miller, Stephen Betts, and Frances McNamara. 2010. "Case Study: Use of Semantic Web Technologies on the BBC Web Sites." Non-profit. *W3C Semantic Web Use Caes and Case Studies*. January. https://www.w3.org/2001/sw/sweo/public/UseCases/BBC/.

Rausas, Matthieu Pelissie du, James Manyika, Eric Hazan, Jacques Bughin, Michael Chui, and Remi Said. 2011. "Internet Matters: The Net's Sweeping Impact on Growth, Jobs, and Prosperity." McKinsey & Company. http://www.mckinsey.com/industries/high-tech/our-insights/internet-matters.

Reagle, Joseph, and Lauren Rhue. 2011. "Gender Bias in Wikiipedia and Britannica." *International Journal of Communication* 5: 1138–58.

Rooney, Joseph Arthur. 1995. "Branding: A Trend for Today and Tomorrow." *Journal of Product & Brand Management* 4 (4): 48–55. http://dx.doi.org/10.1108/10610429510097690.

Ross, Jeanne W., and Peter Weill. 2002. "Six IT Decisions Your IT People Shouldn't Make." *Harvard Business Review*, OnPoint Article, 80 (11): 84–95.

Rowley, Jennifer. 2004. "Online Branding." *Online Information Review* 28 (2): 131–38. http://dx.doi.org/10.1108/14684520410531637.

Rowley, Jennifer, and David Edmundson-Bird. 2013. "Brand Presence in Digital Space:" *Journal of Electronic Commerce in Organizations* 11 (1): 63–78. http://dx.doi.org/10.4018/jeco.2013010104.

Rowntree, Derek. 2004. *Statistics without Tears: A Primer for Non-Mathematicians*. Classic ed. Boston: Pearson/A & B.

RStudio (version 0.99.893). 2016. Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_5) AppleWebKit/601.6.17 (KHTML, like Gecko). En. R Consortium. https://www.rstudio.com/products/rstudio/.

Saha, Saurabh. 2013. "How to Get Verified Name on Google+, Twitter, Facebook and YouTube?" Commercial. *TechGYD - Technology Blog*. June 7.

http://www.techgyd.com/how-to-get-verified-name-on-google-twitter-facebook-and-youtube/3266/.

Sanford, Nevitt. 1970. “Whatever Happened to Action Research?” *Journal of Social Issues* 26 (4): 3–23. http://dx.doi.org/10.1111/j.1540-4560.1970.tb01740.x.

Sauermann, Leo, Richard Cyganiak, and Max Völkel. 2011. “Cool URIs for the Semantic Web.” 07–01. Kaiserslauten; Saarbrücken: Deutsches Forschungszentrum für Künstliche Intelligenz GmbH. http://scidok.sulb.uni-saarland.de/volltexte/2011/3944/pdf/TM_07_01.pdf.

Saunders, Laura. 2015. “Academic Libraries’ Strategic Plans: Top Trends and Under-Recognized Areas.” *The Journal of Academic Librarianship* 41 (3): 285–91. http://dx.doi.org/10.1016/j.acalib.2015.03.011.

Schmachtenberg, Max, Christian Bizer, and Heiko Paulheim. 2014. “State of the LOD Cloud 2014.” University of Mannheim. http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/.

“Search Engine Marketing.” 2016. *Wikipedia: The Free Encyclopedia*. Wikimedia Foundation, Inc. https://en.wikipedia.org/wiki/Search_engine_marketing.

Segal, David. 2011. “Search Optimization and Its Dirty Little Secrets.” *The New York Times*, February 12, sec. Business Day. http://www.nytimes.com/2011/02/13/business/13search.html?_r=1.

“Semantic Triple.” 2016. *Wikipedia: The Free Encyclopedia*. Wikimedia Foundation, Inc. https://en.wikipedia.org/wiki/Semantic_triple.

Sen, Ravi. 2005. “Optimal Search Engine Marketing Strategy.” *International Journal of Electronic Commerce* 10 (1): 17.

Shanks, Justin, and Kenning Arlitsch. 2016. “Making Sense of Researcher Services.” *Journal of Library Administration* 56 (3): 295–316. http://dx.doi.org/10.1080/01930826.2016.1146534.

Singh, Rajesh. 2004. “Branding in Library and Information Context: The Role of Marketing Culture.” *Information Services and Use* 24 (2): 93–98.

Singhal, Amit. 2012. “Introducing the Knowledge Graph: Things, Not Strings.” Corporate. *Google: Official Blog*. May 15. http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html.

Singhal, Amit, and Matt Cutts. 2010. “Using Site Speed in Web Search Ranking.” *Official Google Webmaster Central Blog*. April 9. http://googlewebmastercentral.blogspot.com/2010/04/using-site-speed-in-web-search-ranking.html.

Skiera, Bernd, Jochen Eckert, and Oliver Hinz. 2010. “An Analysis of the Importance of the Long Tail in Search Engine Marketing.” *Electronic Commerce Research and Applications* 9 (6): 488–94. http://dx.doi.org/10.1016/j.elerap.2010.05.001.

Slawski, Bill. 2015. “Google’s Knowledge Cards.” *SEO by the Sea*. March 18. http://www.seobythesea.com/2015/03/googles-knowledge-cards/.

Stinson, Alex, and Jake Orlowitz. 2015. “What Happens When You Give a Wikipedia Editor a Research Library?” Non-profit. *Wikimedia Blog*. March 17. https://blog.wikimedia.org/2015/03/17/wikipedia-research-library/.

Stringer, Ernest T. 2014. *Action Research*. 4th ed. Sage Publications, Inc. https://books.google.com/books?hl=en&lr=&id=nasgAQAAQBAJ&oi=fnd&pg=PR1&dq=stringer+action+research&ots=YSO9c3s4bg&sig=8HVmi-0cz7FG7Mq2wRvG7hREqko#v=onepage&q=stringer%20action%20research&f=false.

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. 2007. “Yago: A Core of Semantic Knowledge.” In *Proceedings of the 16th International Conference on World Wide Web*, 697–706. Banff, Alberta, Canada: ACM Press. http://dx.doi.org/10.1145/1242572.1242667.

———. 2008. “YAGO: A Large Ontology from Wikipedia and WordNet.” *Web Semantics: Science, Services and Agents on the World Wide Web* 6 (3): 203–17. http://dx.doi.org/10.1016/j.websem.2008.06.001.

Suchanek, Fabian, and Gerhard Weikum. 2013. “Knowledge Harvesting in the Big-Data Era.” In , 933. ACM Press. http://dx.doi.org/10.1145/2463676.2463724.

Sullivan, Danny. 2012. “Google Launches Knowledge Graph to Provide Answers, Not Just Links.” *Search Engine Land*. May 16. http://searchengineland.com/google-launches-knowledge-graph-121585.

———. 2014. “The Yahoo Directory - Once the Internet’s Most Important Search Engine - Is to Close.” Commercial. *Search Engine Land*. September 26. http://searchengineland.com/yahoo-directory-close-204370.

Tan, Chun How, Eugene Agichtein, Panos Ipeirotis, and Evgeniy Gabrilovich. 2014. “Trust, but Verify: Predicting Contribution Quality for Knowledge Base Construction and Curation.” In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, 553–62. New York, NY: ACM Press. http://dx.doi.org/10.1145/2556195.2556227.

Tanon, Thomas Pellissier, Denny Vrandecic, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. 2016a. “From Freebase to Wikidata: The Great Migration.” In *Proceedings of the 125th International Conference on World Wide Web*, 1419–28. Montreal: IW3C2. http://dx.doi.org/10.1145/2872427.2874809.

———. 2016b. “From Freebase to Wikidata: The Great Migration.” Google Slides presented at the WWW 2016, Montreal, Canada, April 14. https://docs.google.com/presentation/d/1UtgrHwBLM65T5hmw8udTmgsO9wtphX_09FpQYr09Ie8/edit#slide=id.g6571e2aaf_0_95.

Thalhammer, Andreas, and Achim Rettinger. 2014. “Browsing DBpedia Entities with Summaries.” In *The Semantic Web: ESWC 2014 Satellite Events*, edited by Valentina Presutti, Eva Blomqvist, Raphael Troncy, Harald Sack, Ioannis Papadakis, and Anna Tordai, 8798:511–15. Cham: Springer International Publishing. http://link.springer.com/10.1007/978-3-319-11955-7_76.

The Economist. 1988. “The Year of the Brand,” December 24.

Thornton, Elaine. 2012. “Is Your Academic Library Pinning? Academic Libraries and Pinterest.” *Journal of Web Librarianship* 6 (3): 164–75. http://dx.doi.org/10.1080/19322909.2012.702006.

“Timeline of Web Search Engines.” 2016. *Wikipedia: The Free Encyclopedia*. Wikimedia Foundation, Inc. https://en.wikipedia.org/wiki/Timeline_of_web_search_engines.

Trew, Brandon Kyle, Andrew Swerdlov, and Si-Wai Lai. 2016. Personal knowledge panel interface. USPTO US 9311362 B1, filed March 15, 2013, and issued April 12, 2016.

Uyar, Ahmet, and Farouk Musa Aliyu. 2015. “Evaluating Search Features of Google Knowledge Graph and Bing Satori: Entity Types, List Searches and Query Interfaces.” *Online Information Review* 39 (2): 197–213. http://dx.doi.org/10.1108/OIR-10-2014-0257.

Vaish, Rajan, Keith Wyngarden, Jingshu Chen, Brandon Cheung, and Michael S. Bernstein. 2014. “Twitch Crowdsourcing: Crowd Contributions in Short Bursts of Time.” In

Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 3645–54. Toronto, ON, Canada: ACM Press. http://dx.doi.org/10.1145/2556288.2556996.

Völkel, Max, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. 2006. "Semantic Wikipedia." In *Proceedings of the 15th International Conference on World Wide Web*, 585. Edinburgh, Scotland: ACM Press. http://dx.doi.org/10.1145/1135777.1135863.

Vrandečić, Denny, and Markus Krötzsch. 2014. "Wikidata: A Free Collaborative Knowledgebase." *Communications of the ACM* 57 (10): 78–85. http://dx.doi.org/10.1145/2629489.

Vucovich, Lee A., Valerie S. Gordon, Nicole Mitchell, and Lisa A. Ennis. 2013. "Is the Time and Effort Worth It? One Library's Evaluation of Using Social Networking Tools for Outreach." *Medical Reference Services Quarterly* 32 (1): 12–25. http://dx.doi.org/10.1080/02763869.2013.749107.

Wikidata. 2015. "By Adding to Wikidata, I Have a Free Ticket into Google's Knowledge Graph, Right?" *Help: FAQ/Freebase*. June 18. https://www.wikidata.org/wiki/Help:FAQ/Freebase.

Wikimedia Foundation, Inc. 2016. "Statistics." Special page. *Wikipedia: The Free Encyclopedia*. https://en.wikipedia.org/wiki/Special:Statistics.

"Wikipedia: Contributing to Wikipedia." 2016. *Wikipedia: The Free Encyclopedia*. Wikimedia Foundation, Inc. https://en.wikipedia.org/wiki/Wikipedia:Contributing_to_Wikipedia.

Wilson, T.D. 2000. "Recent Trends in User Studies: Action Resarch and Qualitative Methods." *Information Research* 5 (3). http://www.informationr.net/ir/5-3/paper76.html.

Young, Scott W.H., and Doralyn Rossmann. 2015. "Building Library Community through Social Media." *Information Technology and Libraries* 34 (1): 20.

Appendix A:

Primary Names of ARL Member Libraries.[7]

A

University of Alabama Libraries
University at Albany, SUNY, Libraries
University of Alberta Libraries
University of Arizona Libraries
Arizona State University Libraries
Auburn University Libraries

B

Boston College Libraries
Boston Public Library
Boston University Libraries
Brigham Young University Library
University of British Columbia Library
Brown University Library
University at Buffalo, SUNY, Libraries

C

University of Calgary - Libraries and Cultural Resources
University of California, Berkeley Library
University of California, Davis Library
University of California, Irvine Libraries
UCLA Library
University of California, Riverside Library
UC San Diego Library
University of California, Santa Barbara Libraries
Case Western Reserve University Libraries
Center for Research Libraries
University of Chicago Library
University of Cincinnati Libraries
University of Colorado at Boulder Libraries
Colorado State University Libraries
Columbia University Libraries
University of Connecticut Libraries
Cornell University Library

D

Dartmouth College Library
University of Delaware Library
Duke University Libraries

E

Emory University Libraries

F

University of Florida Libraries
Florida State University Libraries

G

George Washington University Library
Georgetown University Library
University of Georgia Libraries
Georgia Institute of Technology Library
University of Guelph Library

H

Harvard University Libraries
University of Hawai'i at Mānoa Library
University of Houston Libraries
Howard University Libraries

I

University of Illinois at Chicago Library
University of Illinois at Urbana-Champaign Library
Indiana University Libraries Bloomington
University of Iowa Libraries

[7] http://www.arl.org/membership/list-of-arl-members

Iowa State University Library

J
Johns Hopkins University Libraries

K
University of Kansas Libraries
Kent State University Libraries
University of Kentucky Libraries

L
Bibliothèque de l' Université Laval
Library of Congress
Louisiana State University Libraries
University of Louisville Libraries

M
McGill University Library
McMaster University Libraries
University of Manitoba Libraries
University of Maryland Libraries
University of Massachusetts Amherst Libraries
Massachusetts Institute of Technology Libraries
University of Miami Libraries
University of Michigan Library
Michigan State University Libraries
University of Minnesota Libraries
University of Missouri–Columbia Libraries

N
National Agricultural Library
National Archives and Records Administration
National Library of Medicine
National Research Council Canada (NRCC)
University of Nebraska–Lincoln Libraries
University of New Mexico Libraries
New York Public Library
New York State Library
New York University Libraries
University of North Carolina at Chapel Hill Libraries
North Carolina State University Libraries
Northwestern University Library
University of Notre Dame, Hesburgh Libraries

O
Ohio State University Libraries
Ohio University Libraries
University of Oklahoma Libraries
Oklahoma State University Library
University of Oregon Libraries
University of Ottawa Library

P
University of Pennsylvania Libraries
Pennsylvania State University Libraries
University of Pittsburgh Libraries
Princeton University Library
Purdue University Libraries

Q
Queen's University Library

R
Rice University Library
University of Rochester Libraries
Rutgers University Libraries

S
University of Saskatchewan Library
Smithsonian Libraries
University of South Carolina Libraries

University of Southern California Libraries
Southern Illinois University Carbondale Library
Stony Brook University, SUNY, Libraries
Syracuse University Libraries

T
Temple University Libraries
University of Tennessee, Knoxville, Libraries
University of Texas Libraries
Texas A&M University Libraries
Texas Tech University Libraries
University of Toronto Libraries
Tulane University Library

U
University of Utah Library

V
Vanderbilt University Library
University of Virginia Library
Virginia Tech Libraries

W
University of Washington Libraries
Washington State University Libraries
Washington University in St. Louis Libraries
University of Waterloo Library
Wayne State University Libraries
Western University Libraries
University of Wisconsin–Madison Libraries

Y
Yale University Library
York University Libraries

Appendix B: Equations

Equation 1: Results for the pairwise relationship command in R that shows the number of accurate KC that were discovered for either the primary or alternate names of the 125 ARL member libraries = 102. Each of the 125 libraries has a primary name and 94 libraries also have an alternate name, thus the total number KC displayed (102) plus the inaccurate KC (6) plus the KC that failed to display (17) must equal 125. However, this equation does not distinguish whether the KC was found for the primary or the alternate name.

R command string: t(with(SWI,table(PrimOrAltKC,AccurateKCInst)))

Explanation
t = matrix transpose (reverses rows and columns for a table display)
SWI = data frame (the name of the Comma Separated Values (CSV) file)
table = table display

PrimOrAltKC = the spreadsheet column that recorded the display of a KC for either the primary or alternate name of the library. "0" indicates neither primary or alternate name displayed a KC; "1" indicates a primary or alternate name displayed a KC

AccurateKCInst = "0" indicates an inaccurate KC was displayed for the institution, "1" indicates an accurate KC was displayed

Results

```
               PrimOrAltKC
AccurateKCInst   0   1
             0  17   6
             1   0 102
```

Equation 2: Results of the R equation that demonstrates the lack of "same as" comprehension that would allow the search engine to display the same KC regardless of whether the primary or alternate name is searched.

R command string: t(with(SWI,table(AccurateKCInst,SameAs)))

Explanation of terms
AccurateKCInst = 0 indicates an inaccurate KC, 1 indicates an accurate KC
SameAs = 0 indicates a different KC displayed for the alternate and primary names, 1 indicates the same KC displayed

Results

```
       AccurateKCInst
SameAs  0  1
     0 23 56
     1  0 46
```

Equation 3: Results of the R equation that demonstrates the number of accurate KC that displayed for primary and alternate ARL library names.

```
t(with(SWI,table(Primary,AccurateKC)))

         Primary
AccurateKC  0  1
         0 20 67
         1 74 58
```

Equation 4: Various pairwise relationship equations that show if a record or article exists in each knowledge base for the primary and alternate names of the ARL libraries

```
t(with(SWI,table(GMB,Primary)))
       GMB
Primary  0  1
      0 54 40
      1 97 28

t(with(SWI,table(Gplus,Primary)))
       Gplus
Primary  0  1  2
      0 58 17 19
      1 78 25 22

t(with(SWI,table(Wikipedia,Primary)))
       Wikipedia
Primary  0  1
      0 52 42
      1 85 40

t(with(SWI,table(WikipediaInfobox,Primary)))
       WikipediaInfobox
Primary  0  1  2
      0 52 16 26
      1 85 10 30

t(with(SWI,table(DBpedia,Primary)))
       DBpedia
Primary  0  1
      0 55 39
      1 95 30

t(with(SWI,table(Wikidata,Primary)))
       Wikidata
Primary  0  1
      0 57 37
      1 99 26
```

Equation 5: R command string and results showing three-way relationship between Primary/Alternate names, GMB profiles, and Accurate KC

```
(with(SWI,table(Primary,GMB,AccurateKC)))

, , AccurateKC = 0

       GMB
Primary  0  1
      0 19  1
      1 62  5


, , AccurateKC = 1

       GMB
Primary  0  1
      0 35 39
      1 35 23
```

Equation 6: R command string and resulting table showing four-way relationship between accurate KC, Wikipedia articles, and Descriptions in the KC

```
R command string:
t(with(SWI,table(Wikipedia,Description,AccurateKC)))

Results
, AccurateKC = 0

         Description
Wikipedia  0  1
        0 66  1
        1 18  2

, , AccurateKC = 1

         Description
Wikipedia  0  1
        0 60 10
        1 15 47
```

Equation 7: Logistic regression command string and explanation of components to predict odds of independent variables affecting the presence of the Description group in the KC

R command string:

```
fit_d<-
glm(factor(Description)~GMB+Wikipedia+Wikidata,data=subset(SWI,Accurate=="1"),f
amily="binomial")
summary(fit_d)
exp(coef(fit_d))
exp(confint(fit_d))
exp(cbind(OR = coef(fit_d), confint(fit_d)))
```

Explanation of terms:

fit_d – establishes the group for the Description element and names it "d"
glm – generalized linear model
factor – turns outcome variable into a categorical variable
Description – outcome variable name
GMB+Wikipedia+Wikidata – these are the independent variables used in the model, against which the outcome variables were compared
SWI – represents the data frame
Accurate=="1" – indicates that an accurate KC must exist to be counted in this calculation
family – indicates which generalized linear model is desired, which in this case is the difference between two terms (binomial)

summary – asks R to display output of results of the model
exp(coef) – exponentiates the coefficients and interprets them as odds-ratios.
exp(confint(fit_d)) – shows confidence intervals
exp(cbind) – binds the coefficients and confidence intervals into a single table

Equation 8: Logistic regression command string used to predict odds of independent variables affecting the presence of the Appearance group in accurate KC

R command string

```
fit_a<-
glm(factor(Appearance)~GMB+Wikipedia+Wikidata,data=subset(
SWI,Accurate=="1"),family="binomial")
```

Equation 9: Logistic regression command string to predict odds of independent variables affecting the presence of the Contact group in accurate KC

R command string

```
fit_c<-
glm(factor(Contact)~GMB+Wikipedia+Wikidata,data=subset(SWI,Accurate=="1"),famil
y="binomial")
```

Appendix C: Case Studies

Montana State University Library

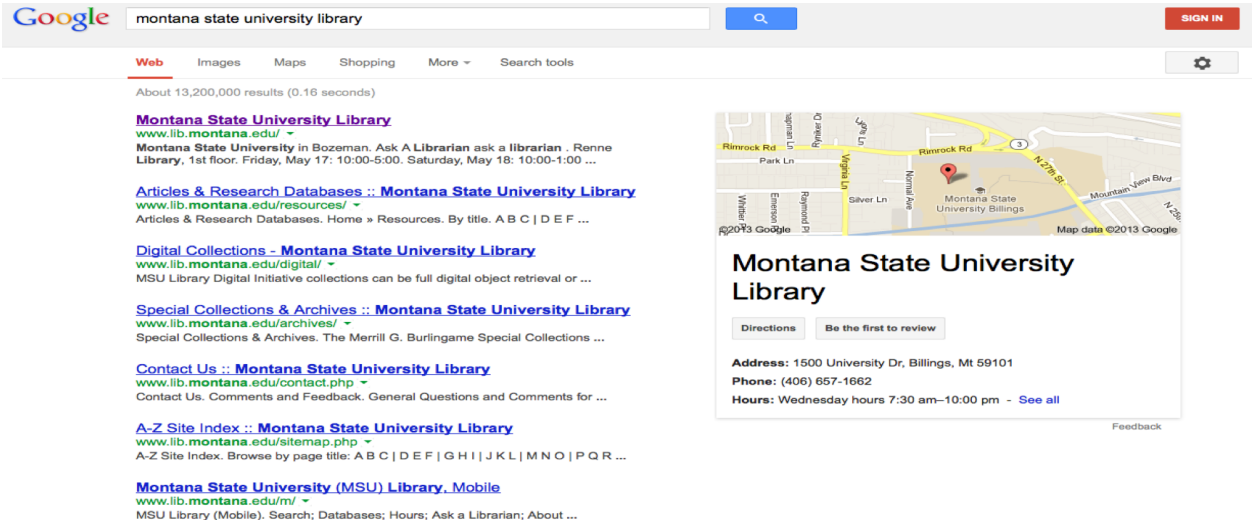

Figure 44: An inaccurate KC displayed for Montana State University Library as late as May 15, 2013

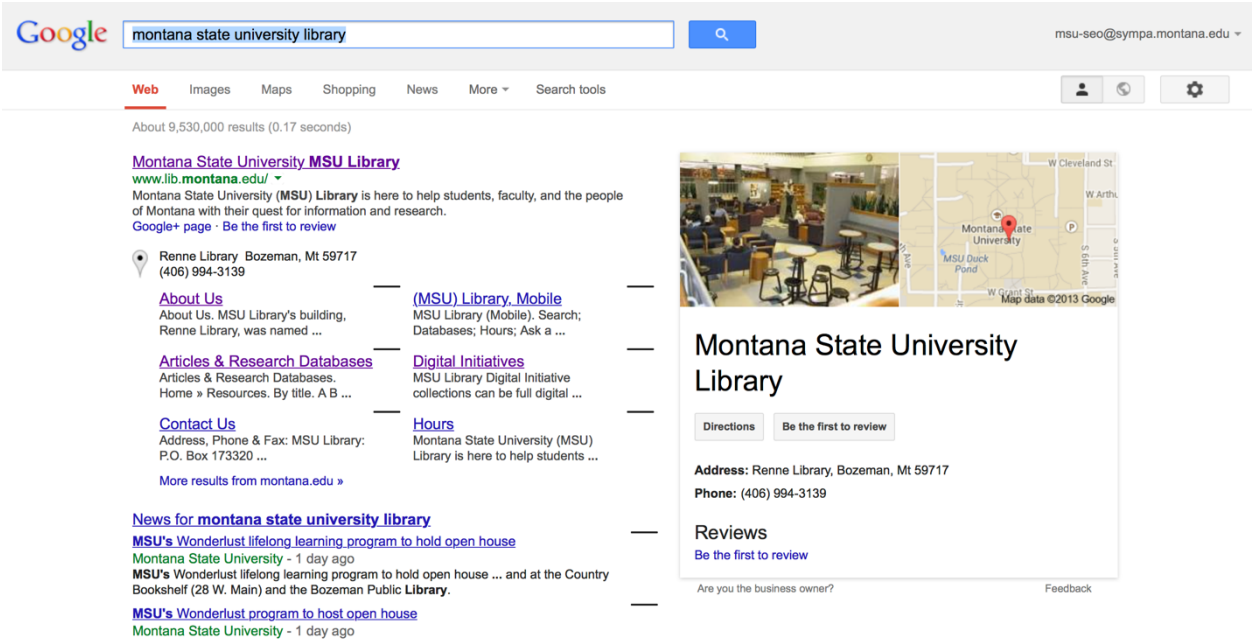

Figure 45: First appearance of an accurate KC for Montana State University Library on September 5, 2013.

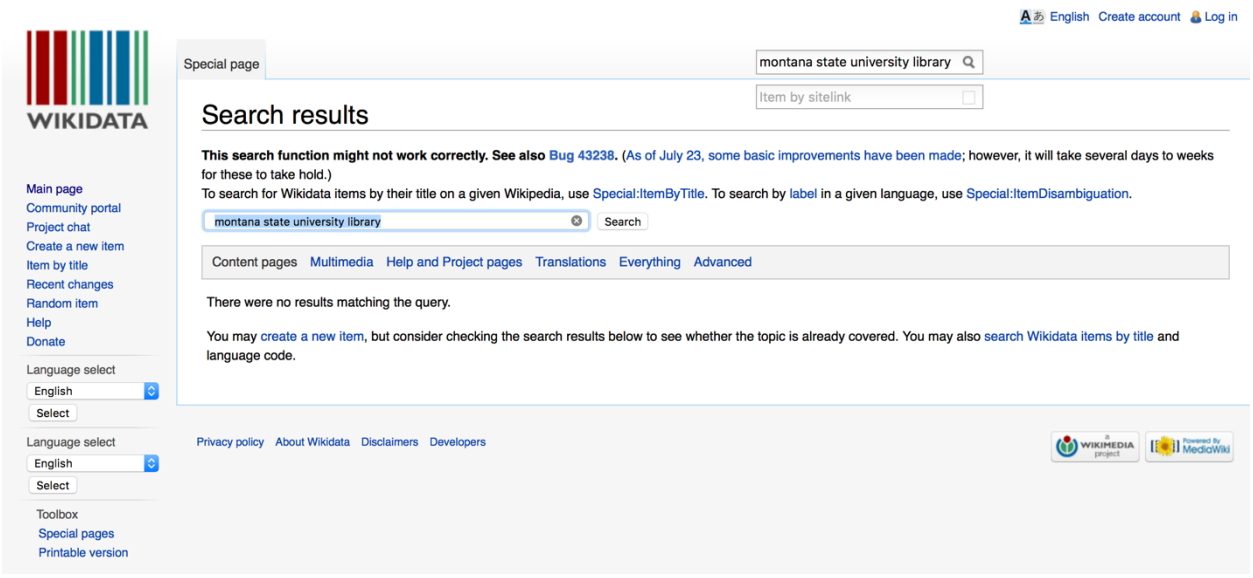

Figure 46: Wikidata lacked a record for the Montana State University Library as late as September 26, 2013

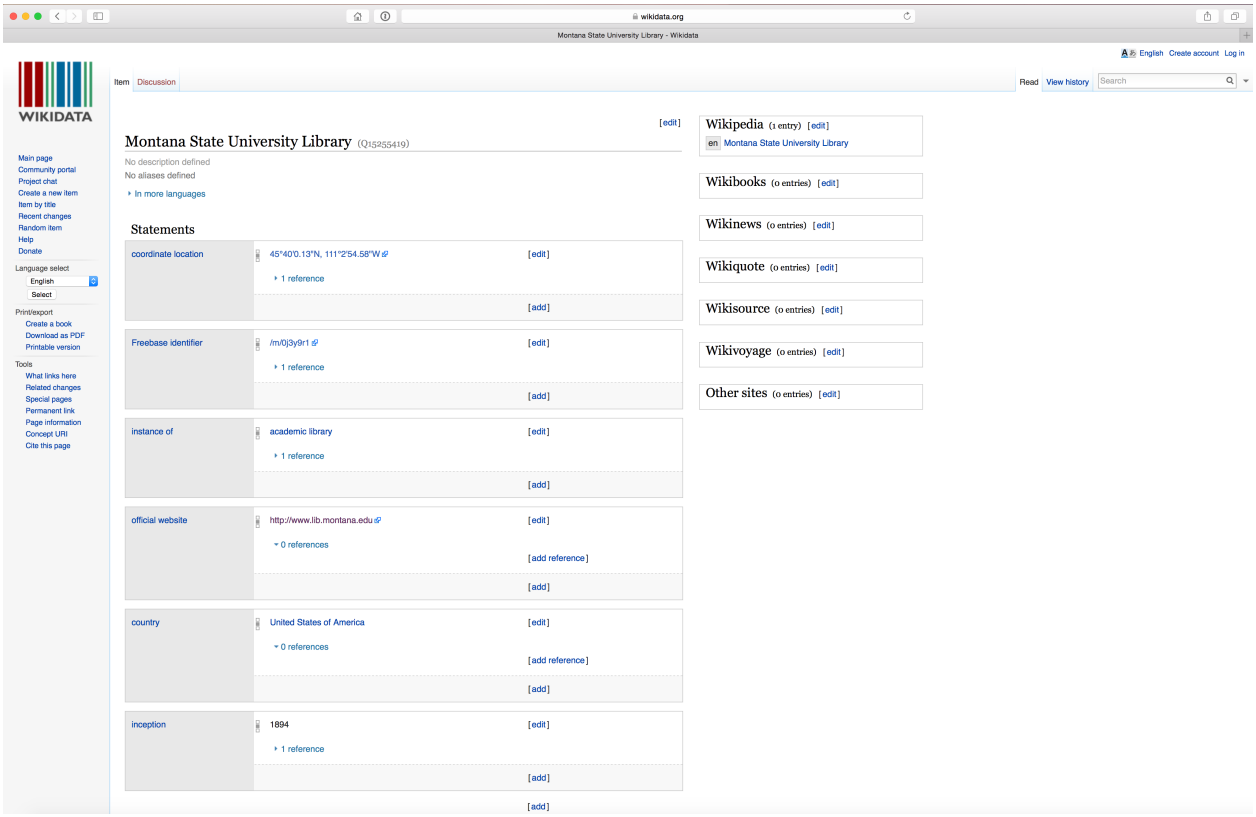

Figure 47: A Wikidata record was evident for the MSU Library on June 26, 2015

McMaster University Library

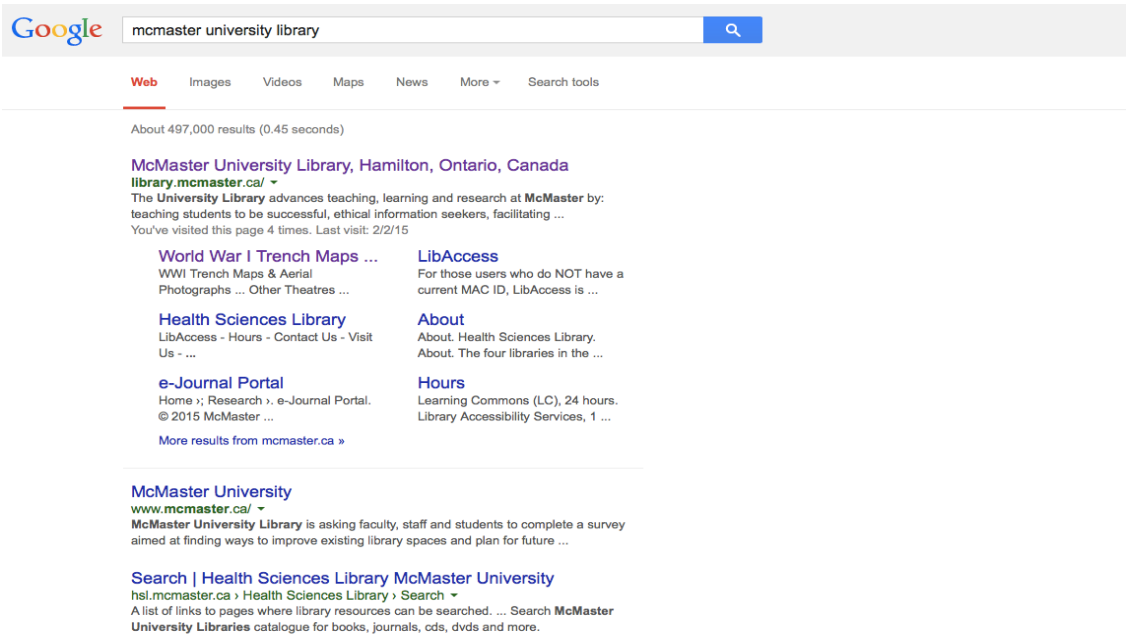

Figure 48: No KC existed for McMaster University Library on February 19, 2015

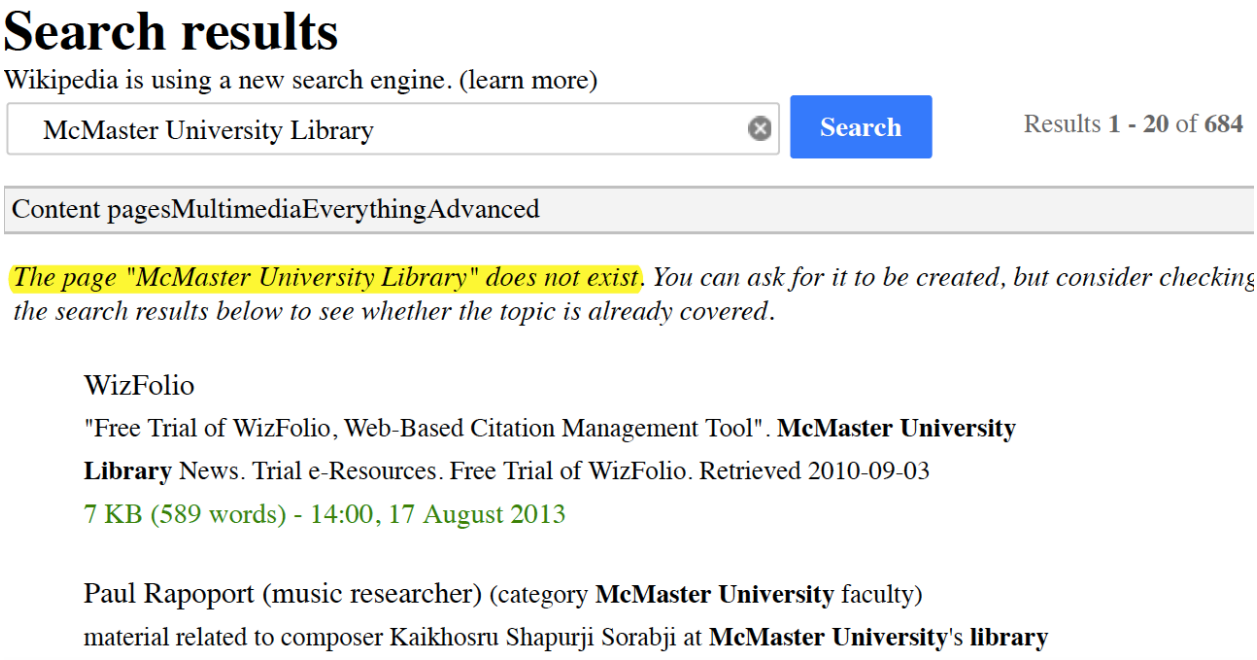

Figure 49: Wikipedia lacked an article for McMaster University Library on December 21, 2014

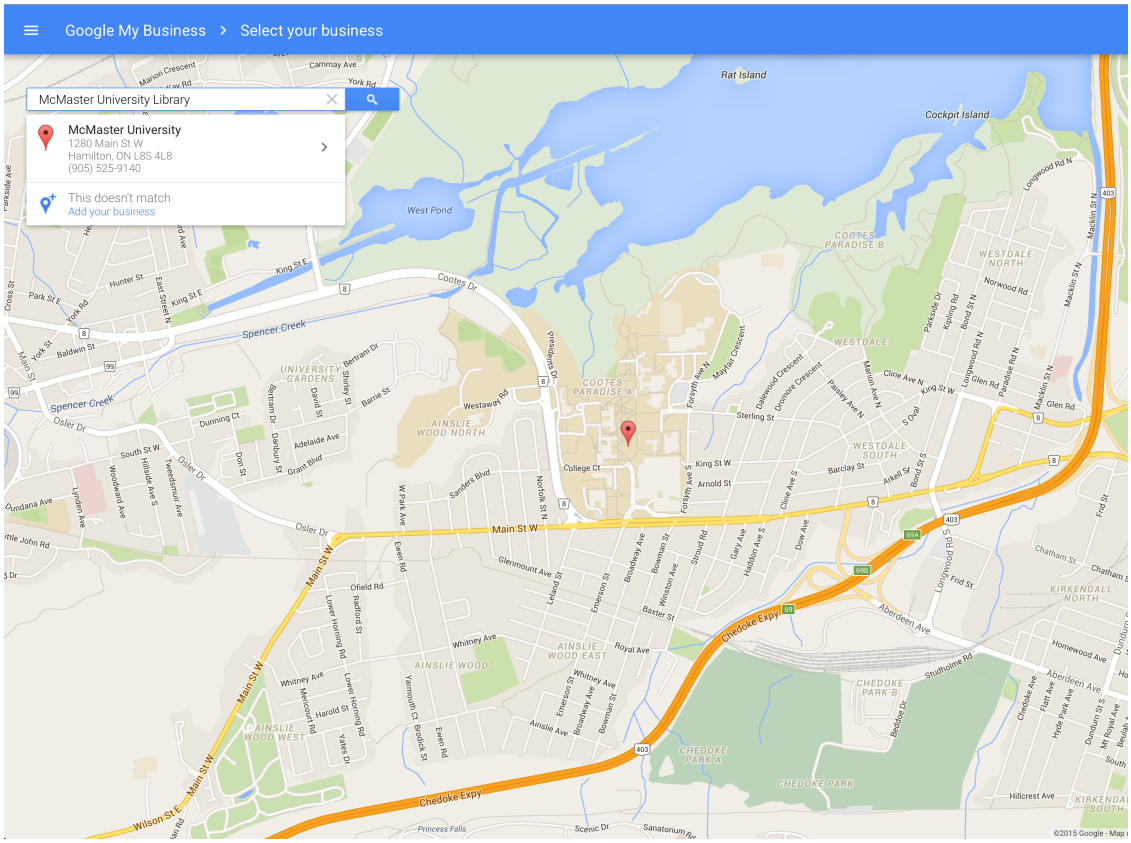

Figure 50: GMB lacked a claimed and verified business profile for McMaster University Library on December 6, 2015

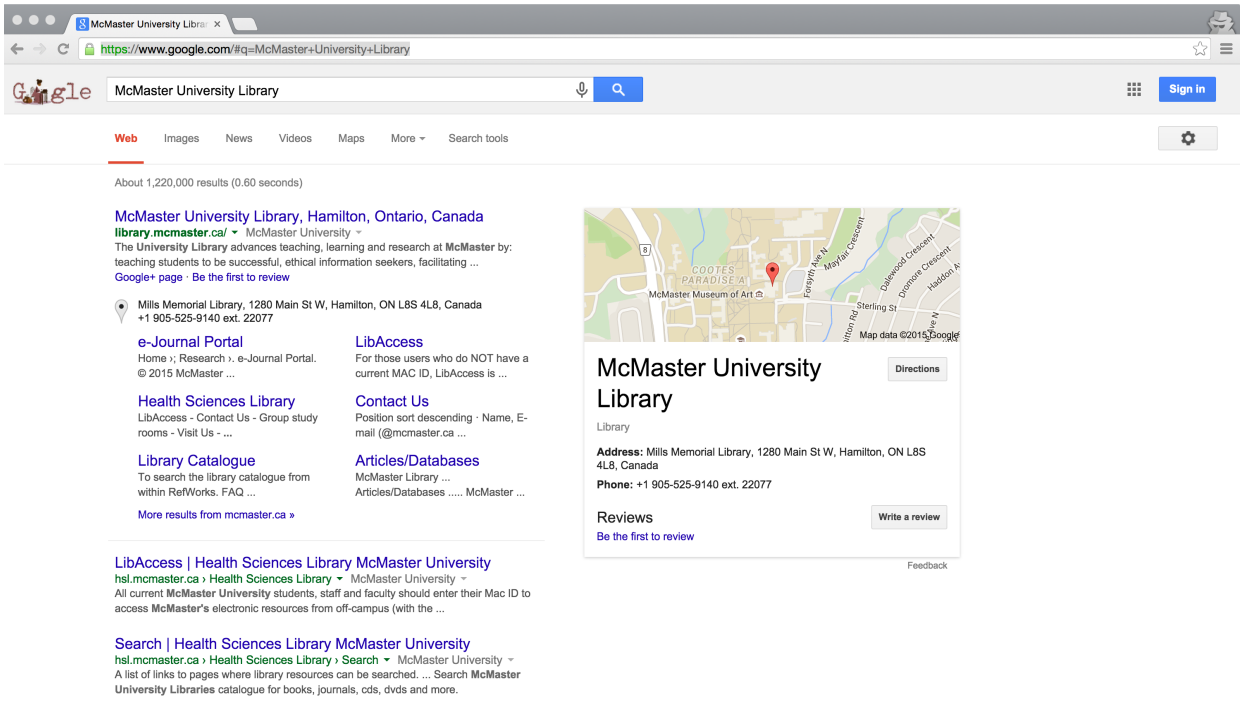

Figure 51: The beginnings of a KC (lacking a description) for McMaster University Library on July 16, 2015

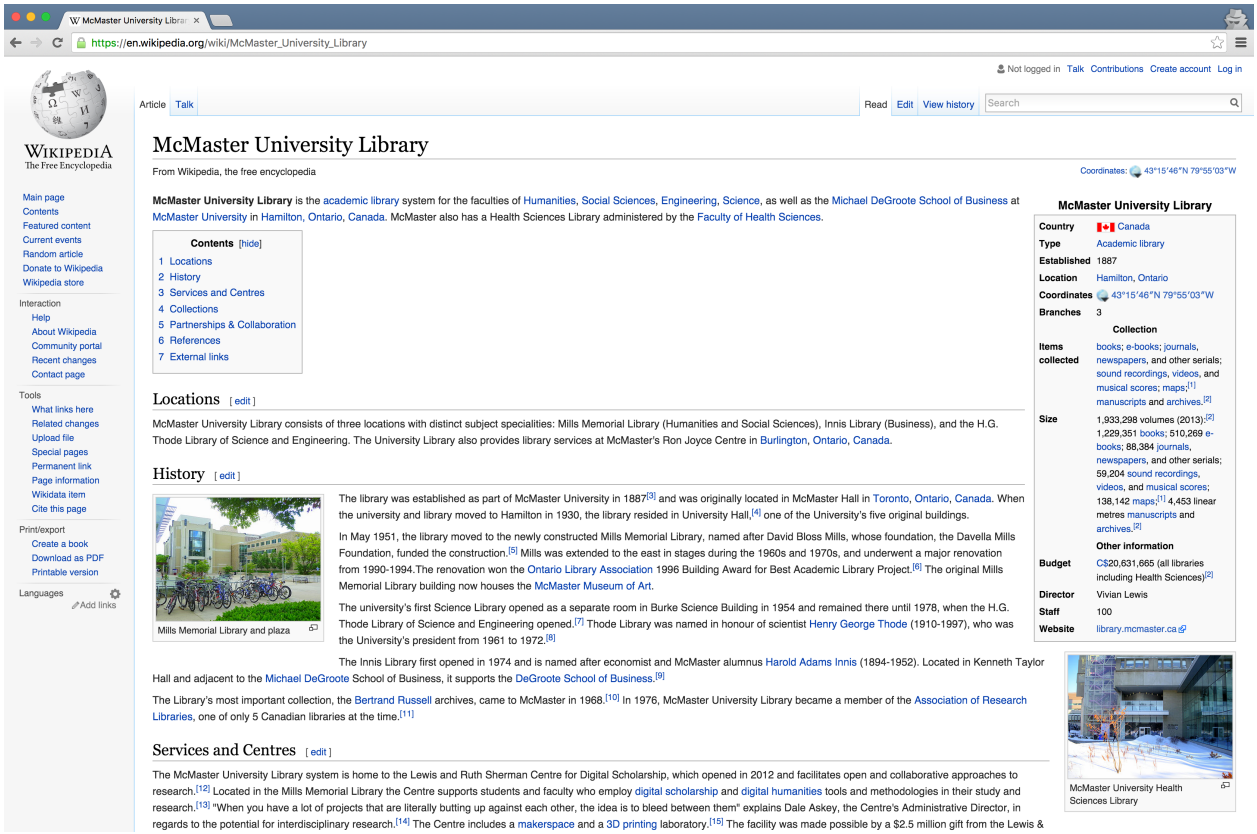

Figure 52: Wikipedia article for McMaster University Library captured on January 3, 2016

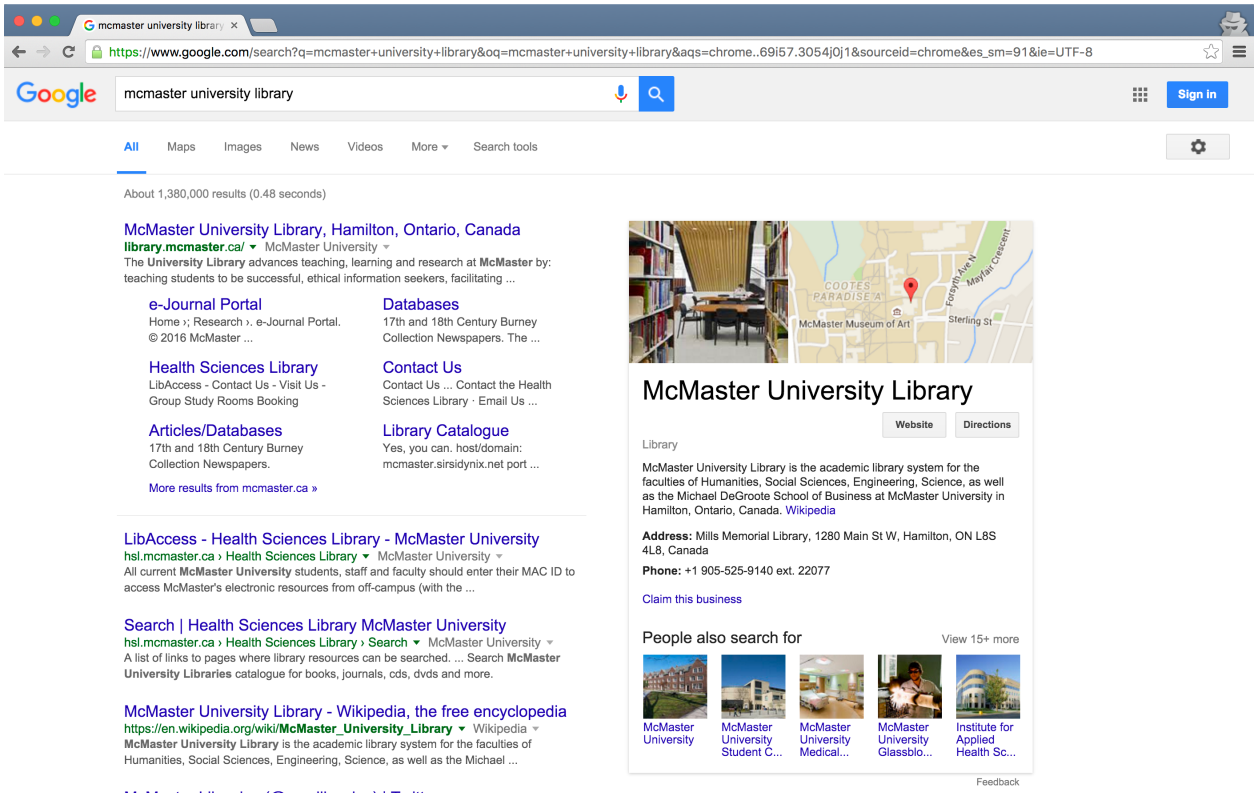

Figure 53: Accurate KC, with description, for McMaster University Library on February 10, 2016

Coalition for Networked Information

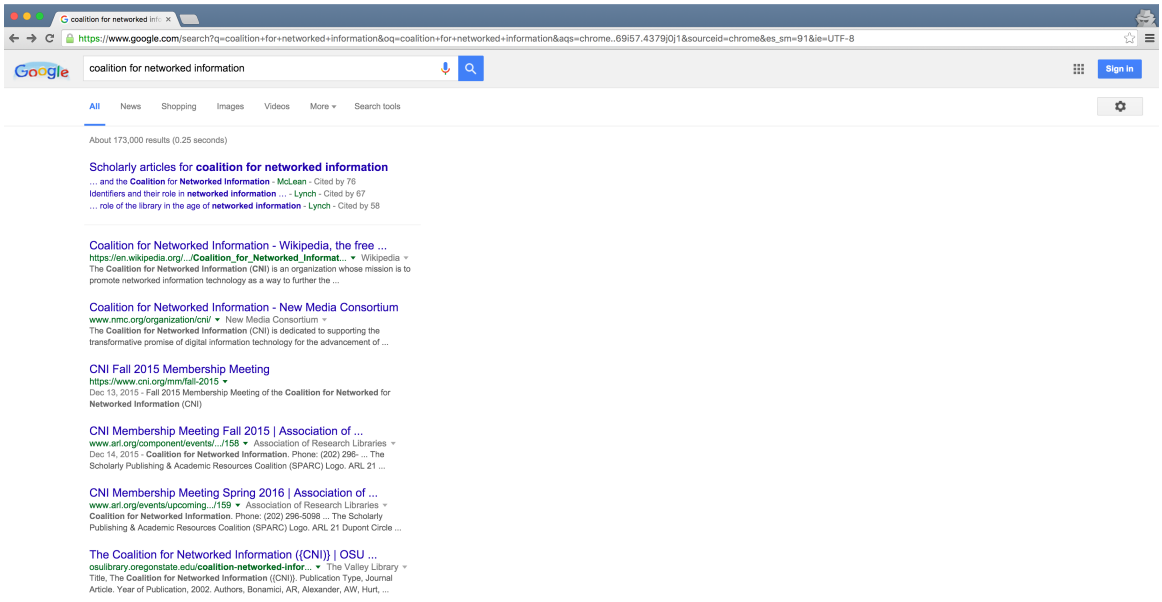

Figure 54: No KC in evidence for CNI in Google SERP on December 22, 2015

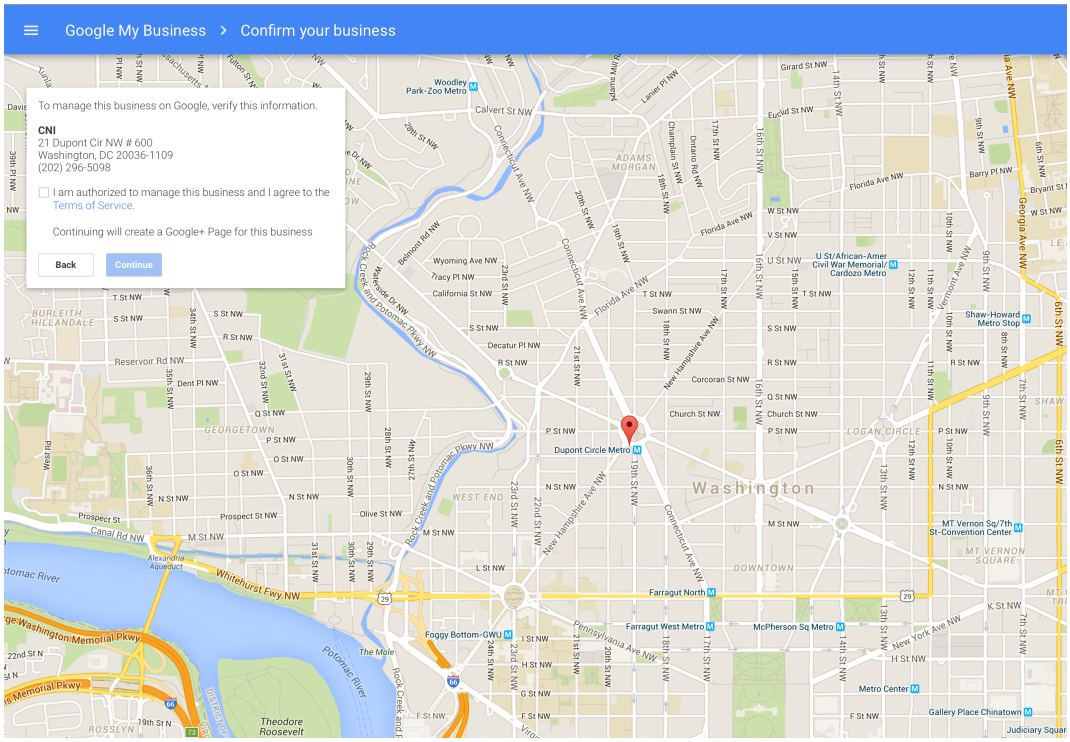

Figure 55: CNI business had not been claimed in GMB as of December 6, 2015

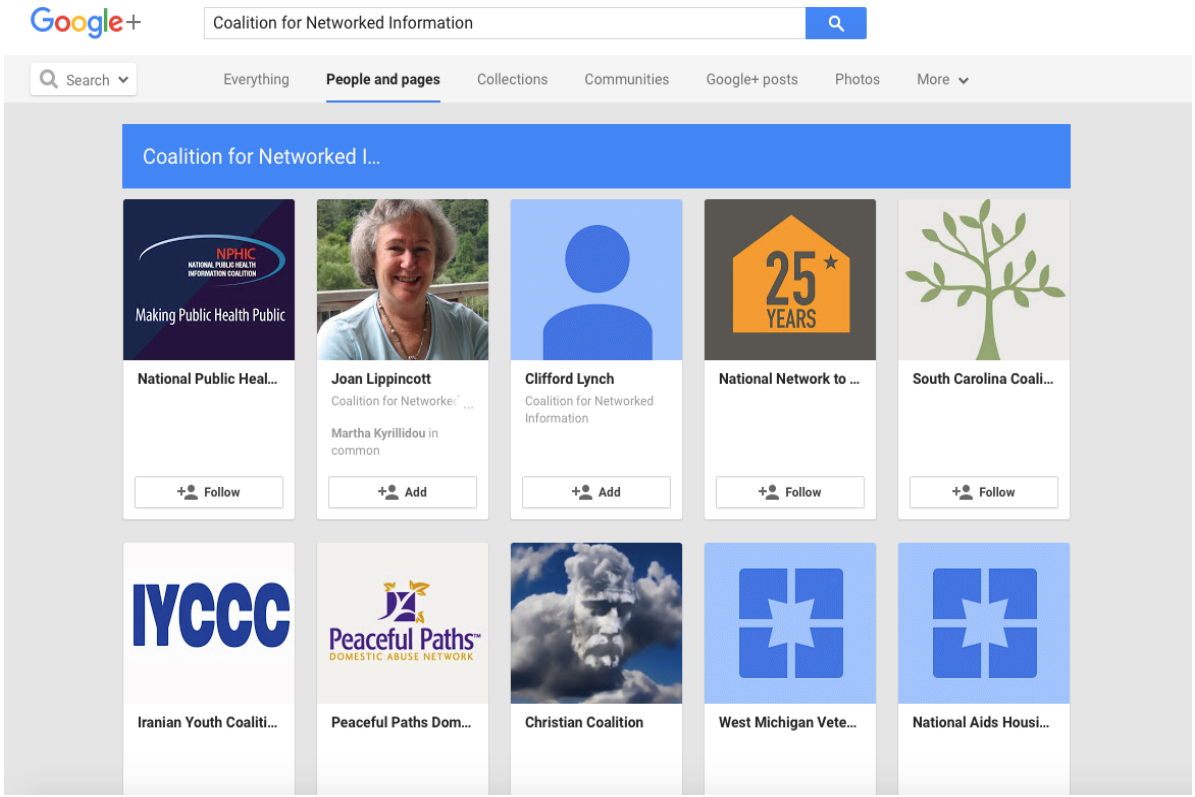

Figure 56: CNI lacked a Google+ profile on October 30, 2015

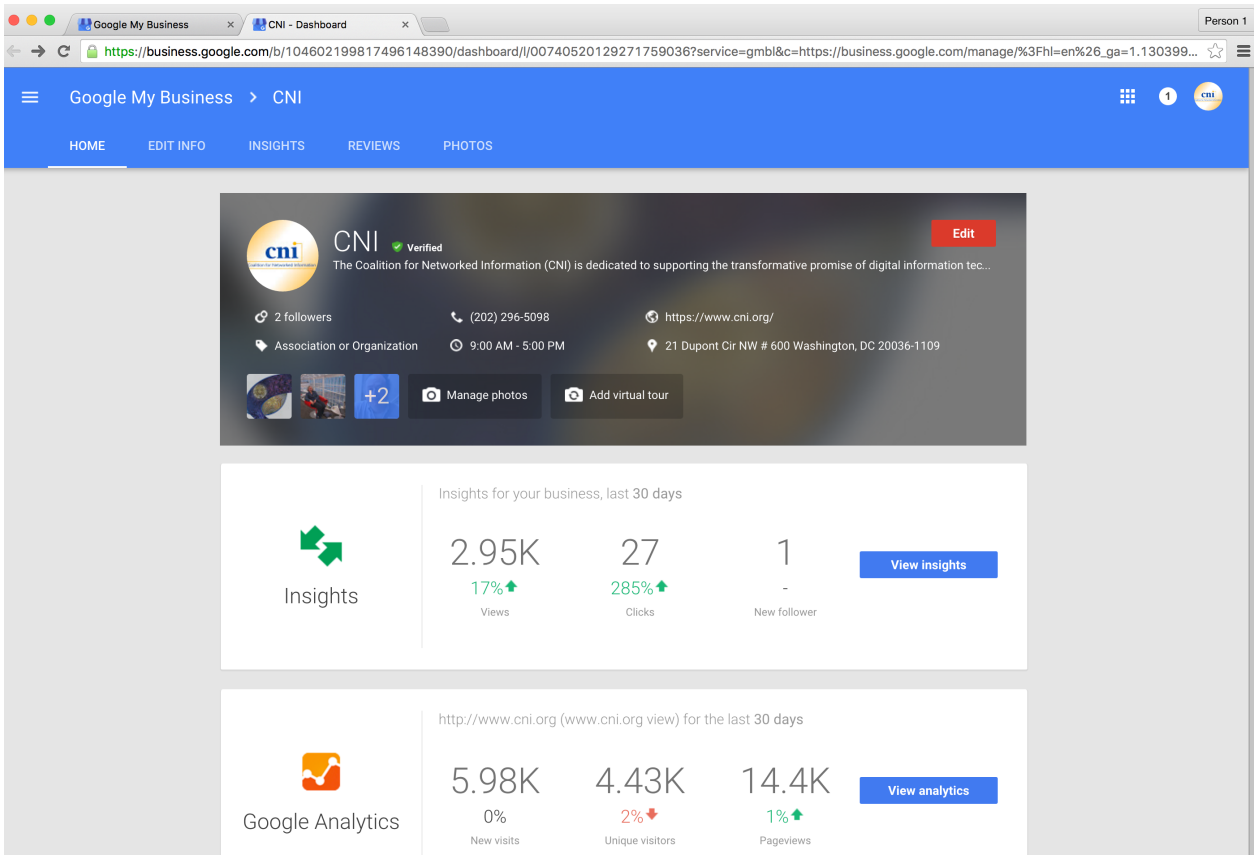

Figure 57: GMB showing claimed and verified profile for CNI on March 10, 2016

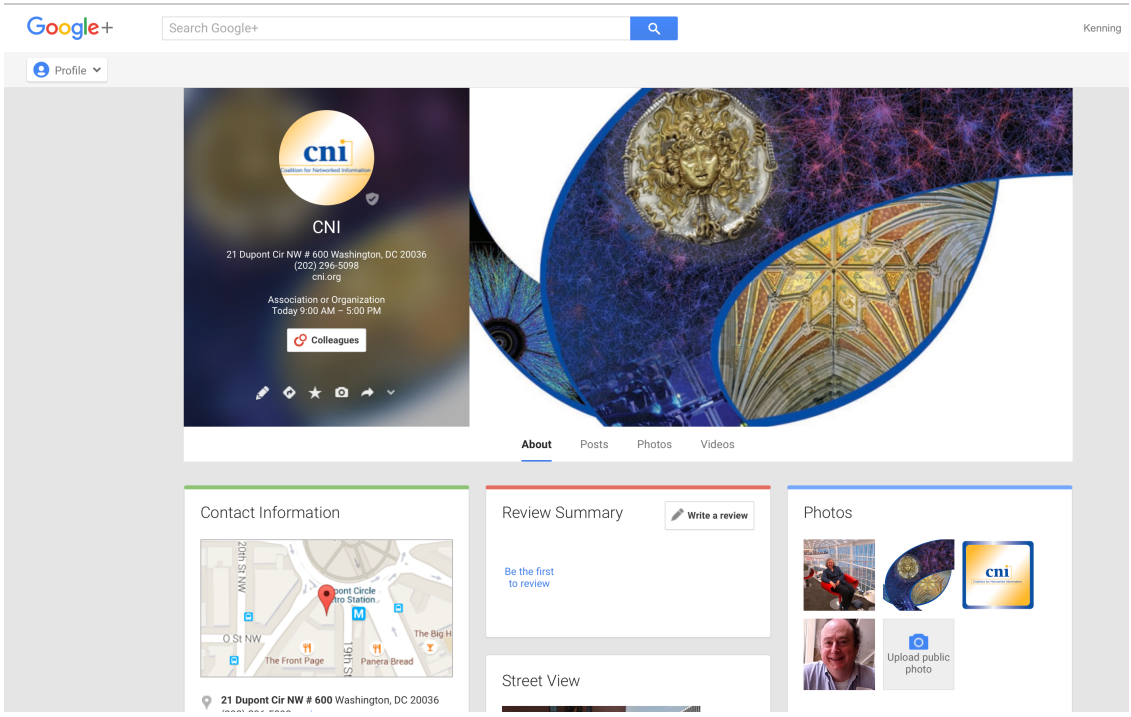

Figure 58: Google+ showing verified profile for CNI on January 6, 2016

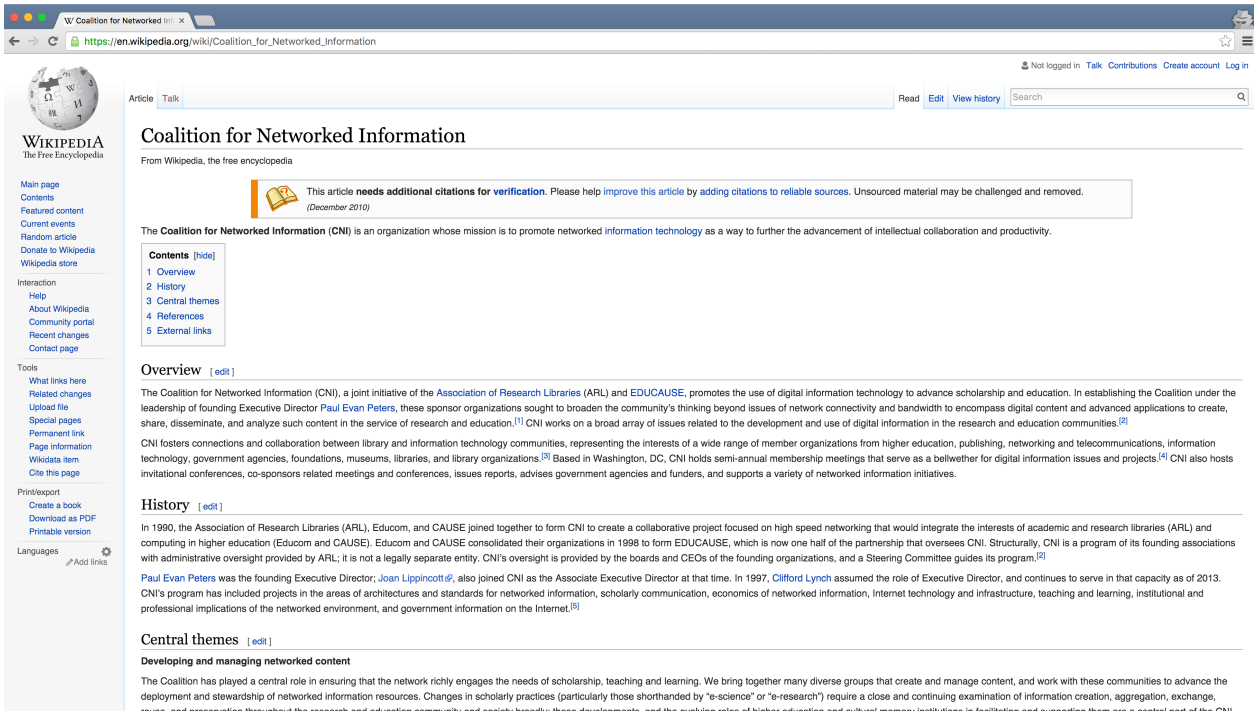

Figure 59: Wikipedia showing flagged article for CNI and lacking infobox on December 22, 2015

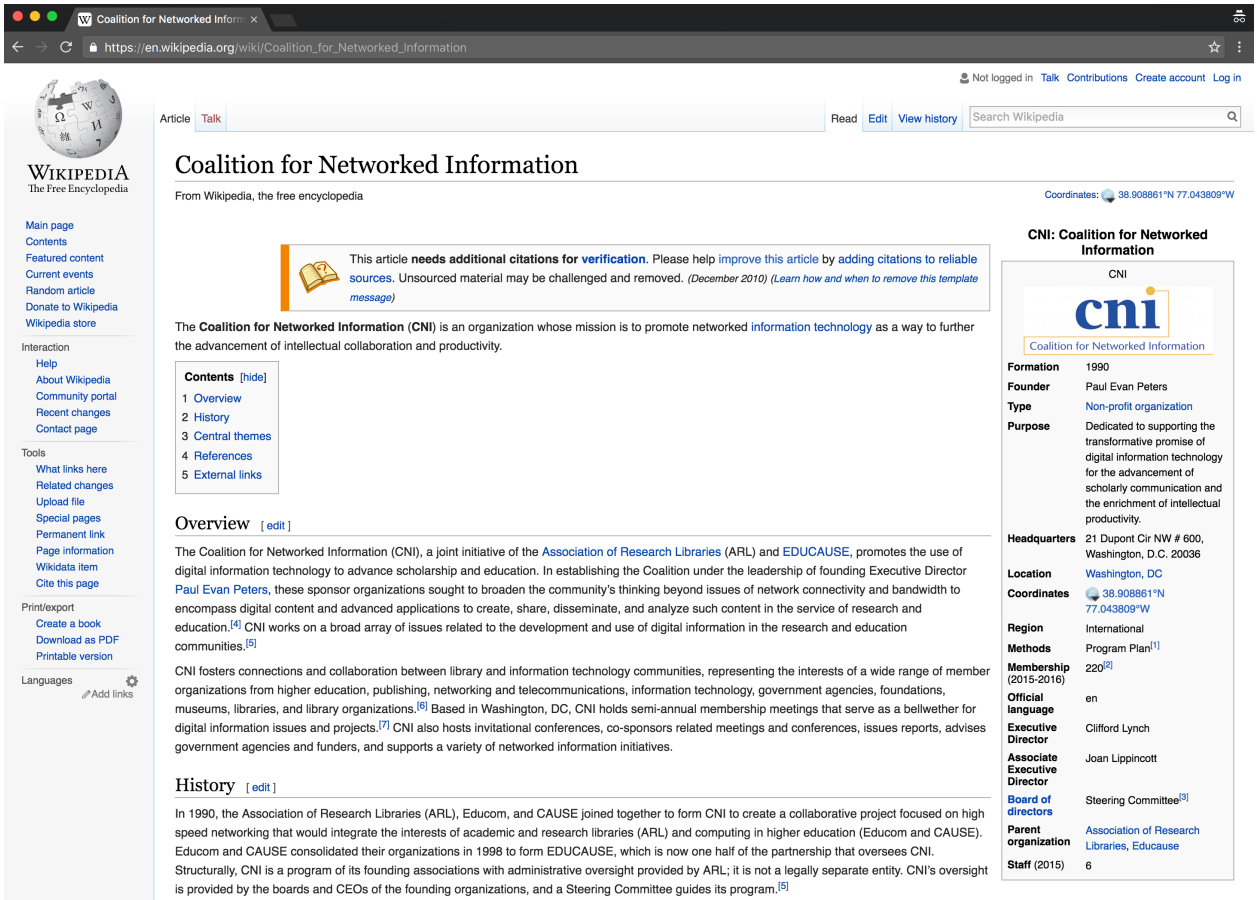

Figure 60: Wikipedia showing article with infobox for CNI on November 22, 2016

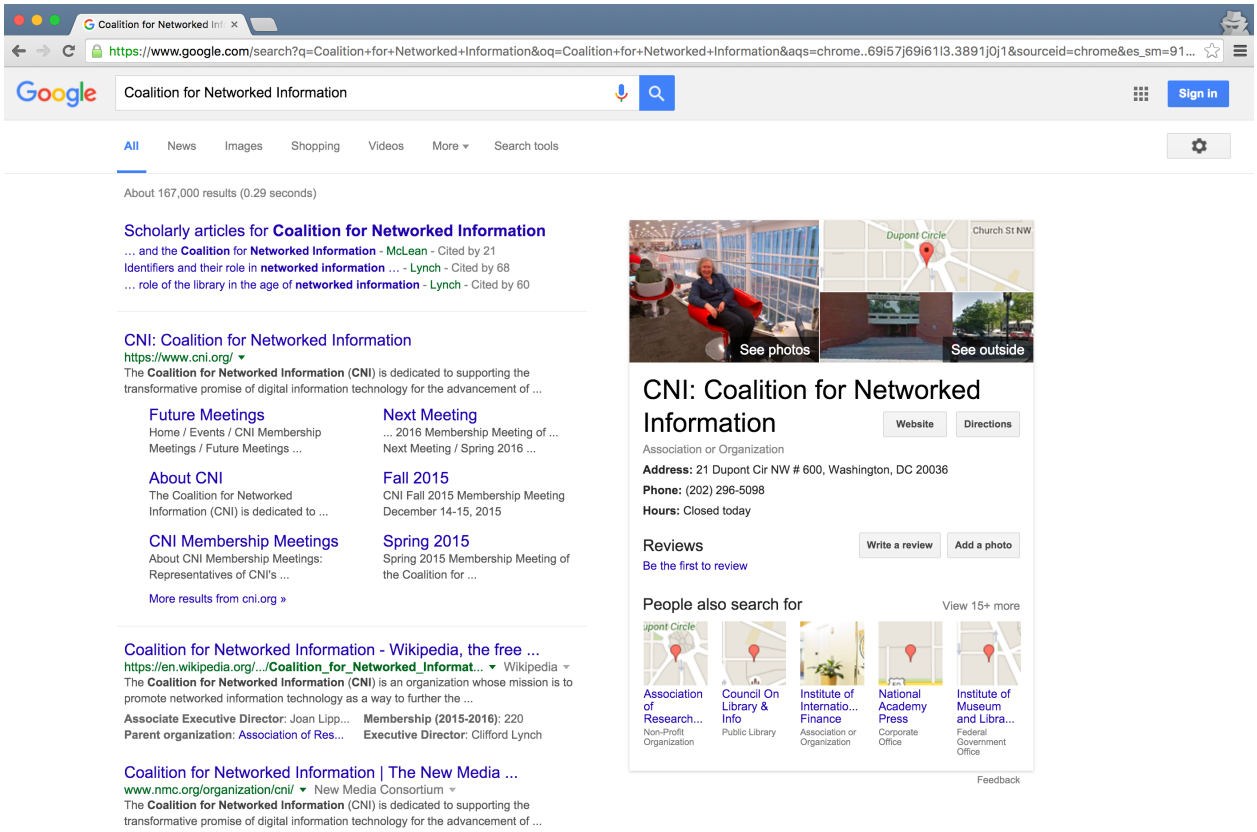

Figure 61: KC showing in Google SERP for CNI on March 12, 2016

Appendix D: MSU Academic Organizations

This appendix shows screen capture images to support the data in Table 9 in Chapter 6. Screen capture files were collected toward the end of December 2015.

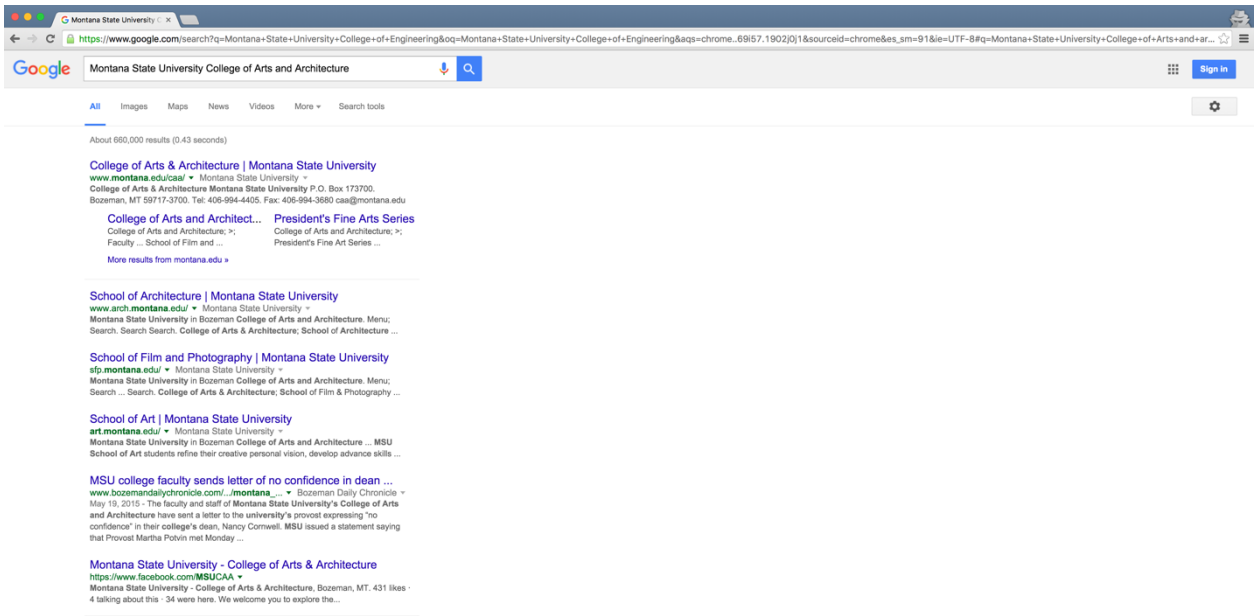

Figure 62: MSU College of Arts and Architecture missing a KC

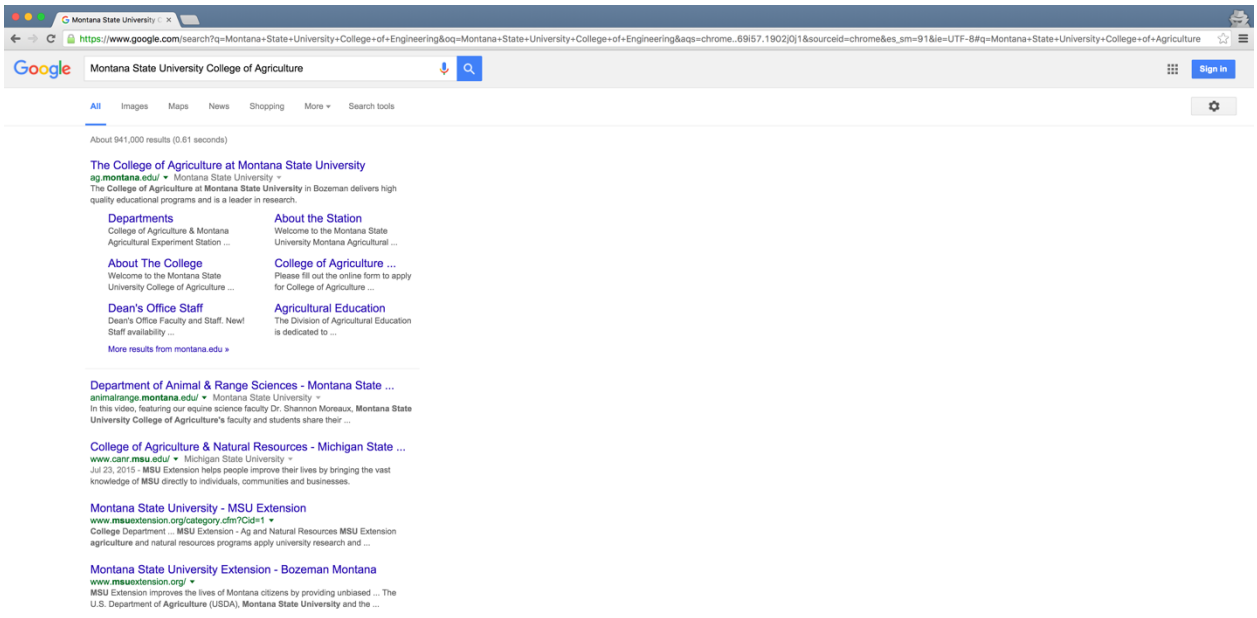

Figure 63: MSU College of Agriculture missing a KC

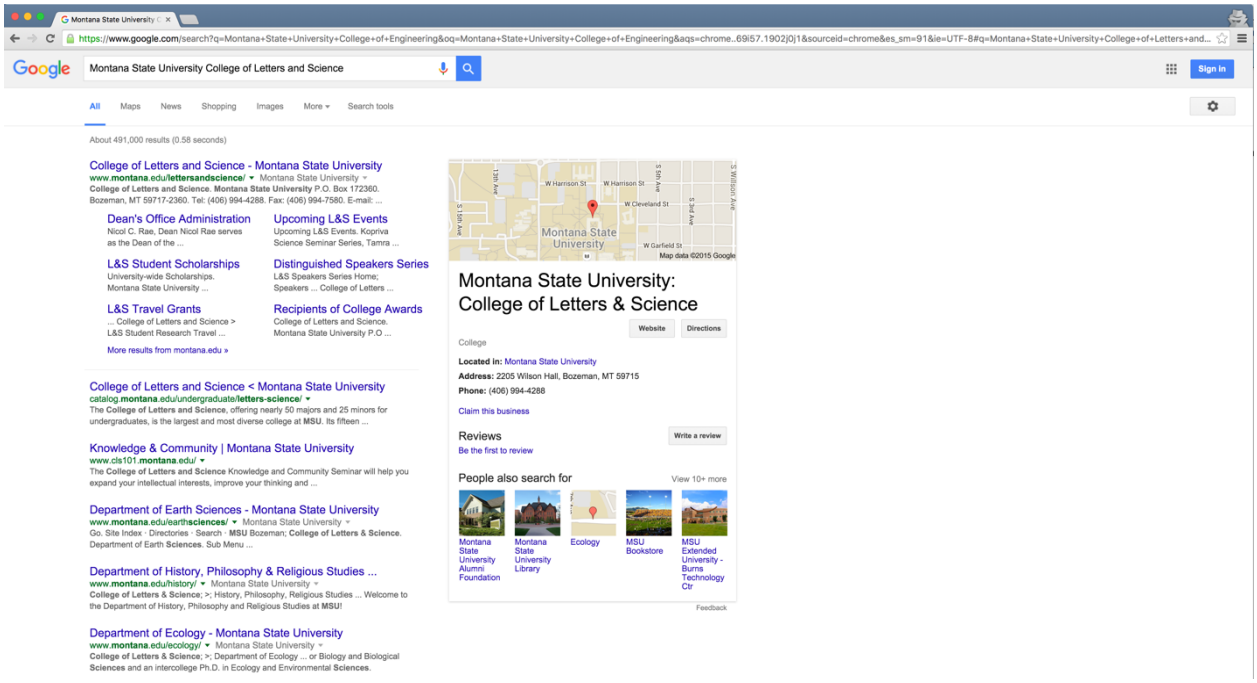

Figure 64: MSU College of Letters and Science showing a minimal KC indicating an unclaimed business

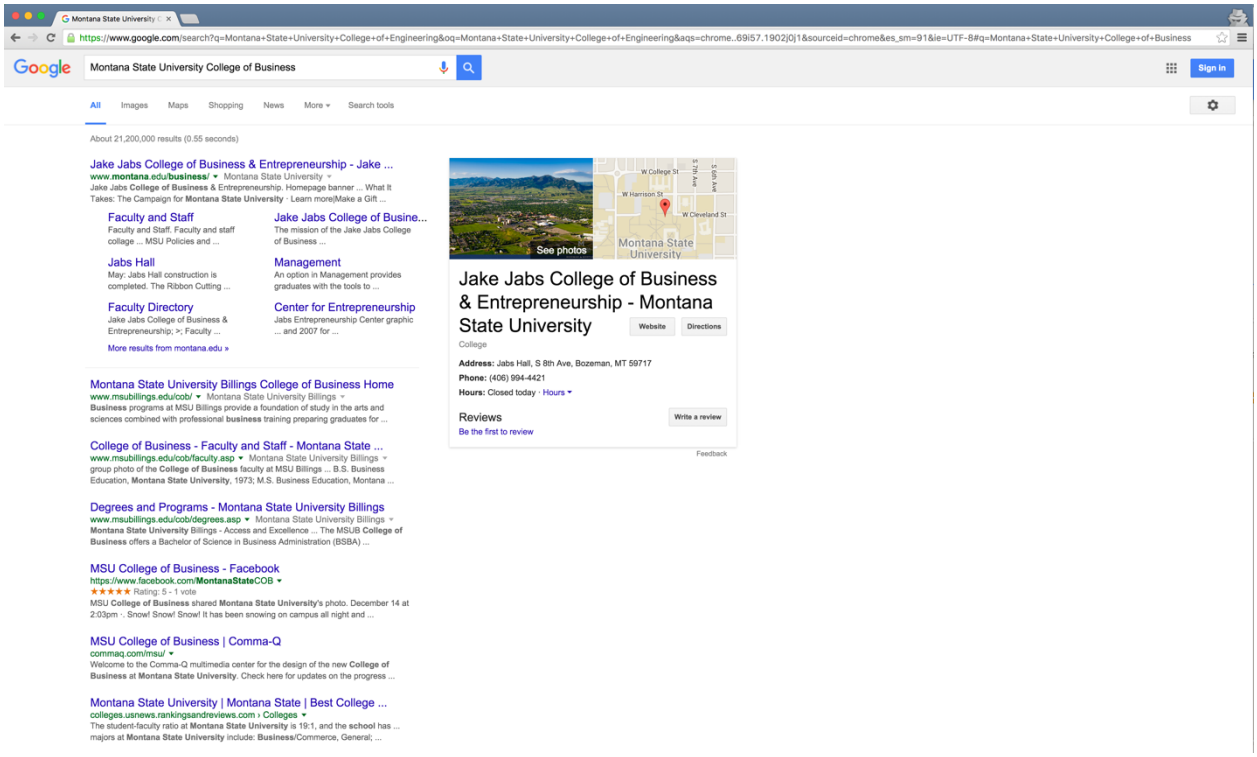

Figure 65: MSU College of Business showing a small KC that resulted from recent intervention by the MSU Library

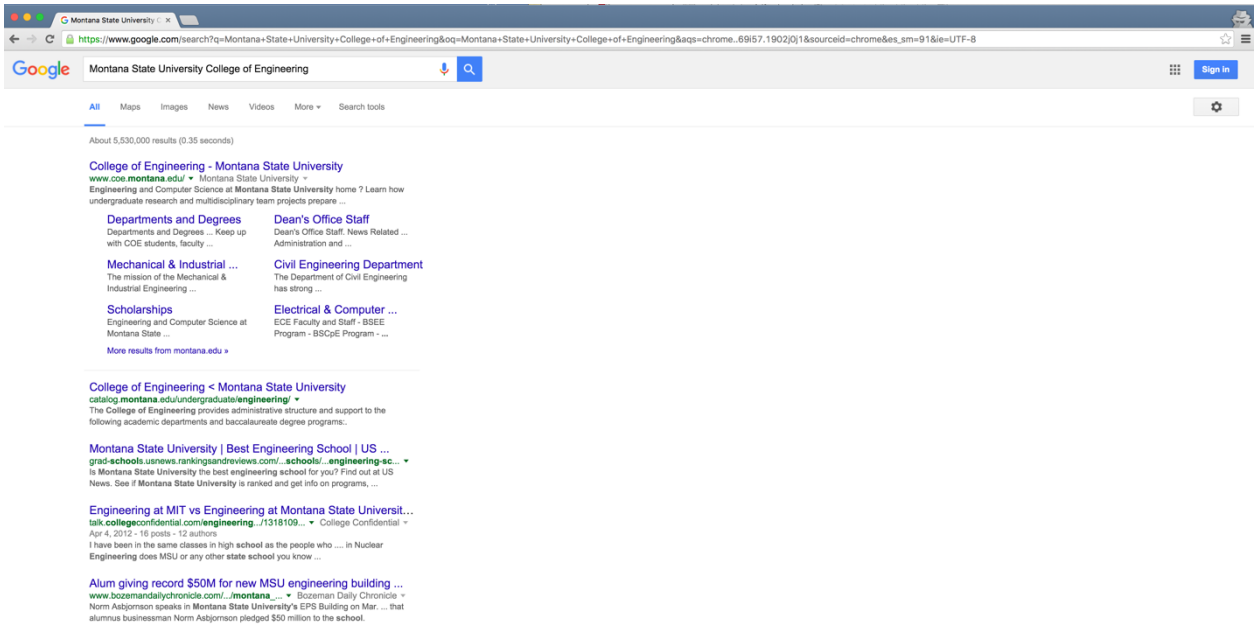

Figure 66: MSU College of Engineering lacking a KC

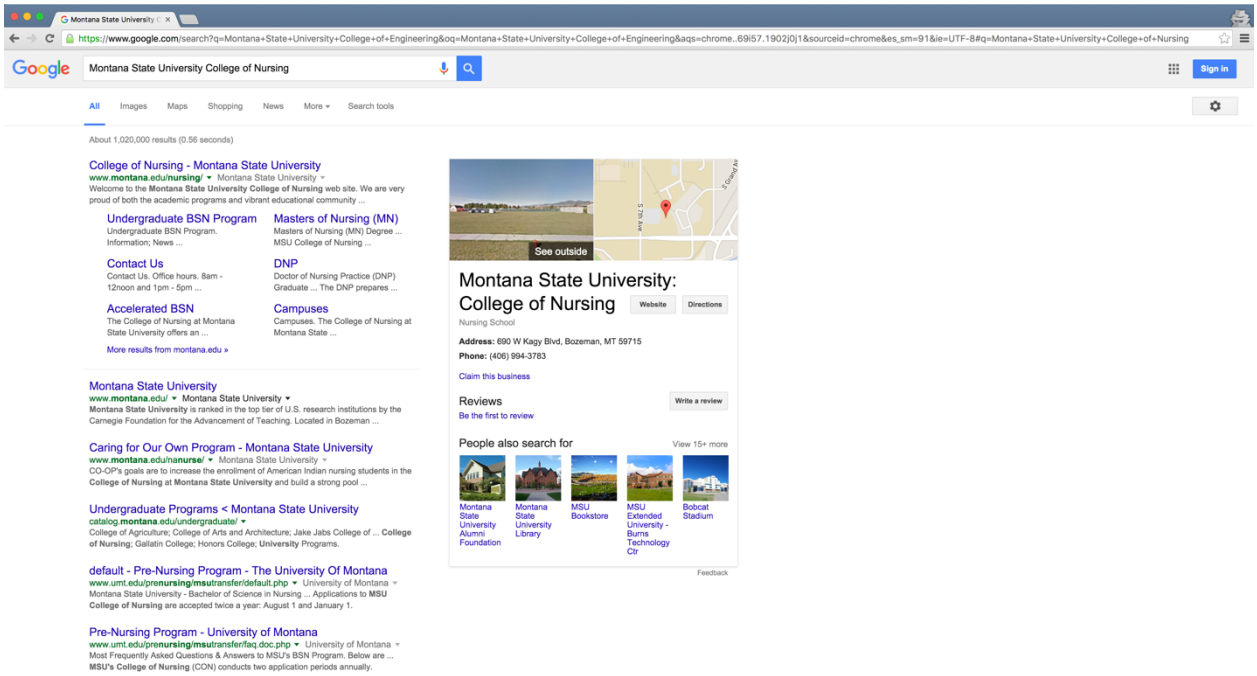

Figure 67: MSU College of Nursing KC with an inaccurate address

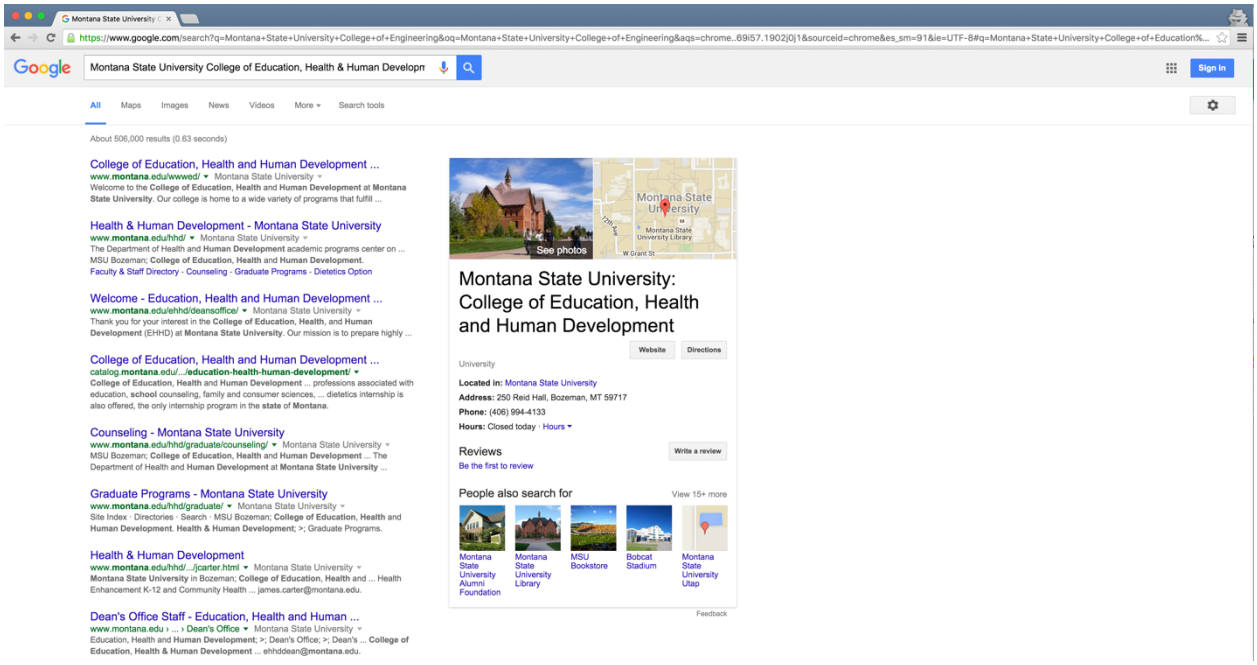

Figure 68: MSU College of Education, Health, and Human Development showing KC that lacks a description.

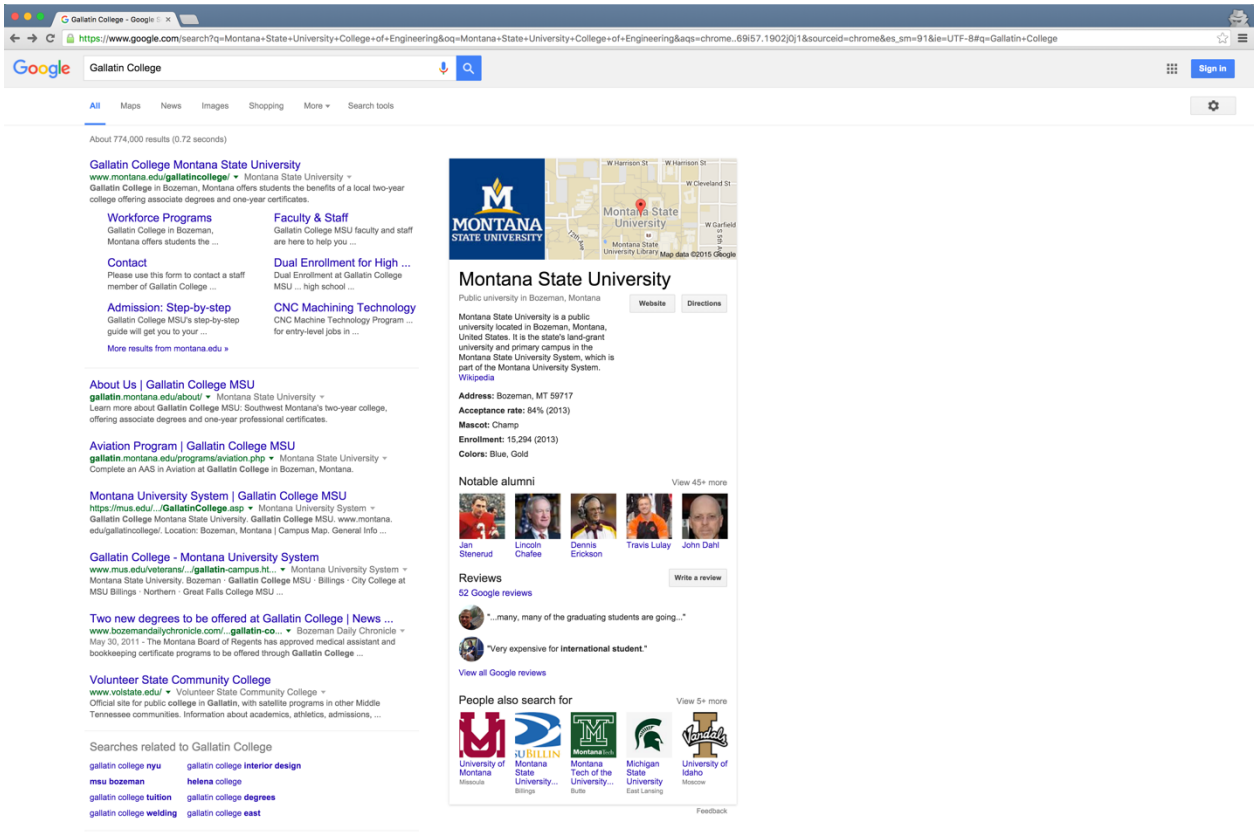

Figure 69: MSU Gallatin College showing KC for the parent institution (this would be considered an inaccurate KC for the search conducted)

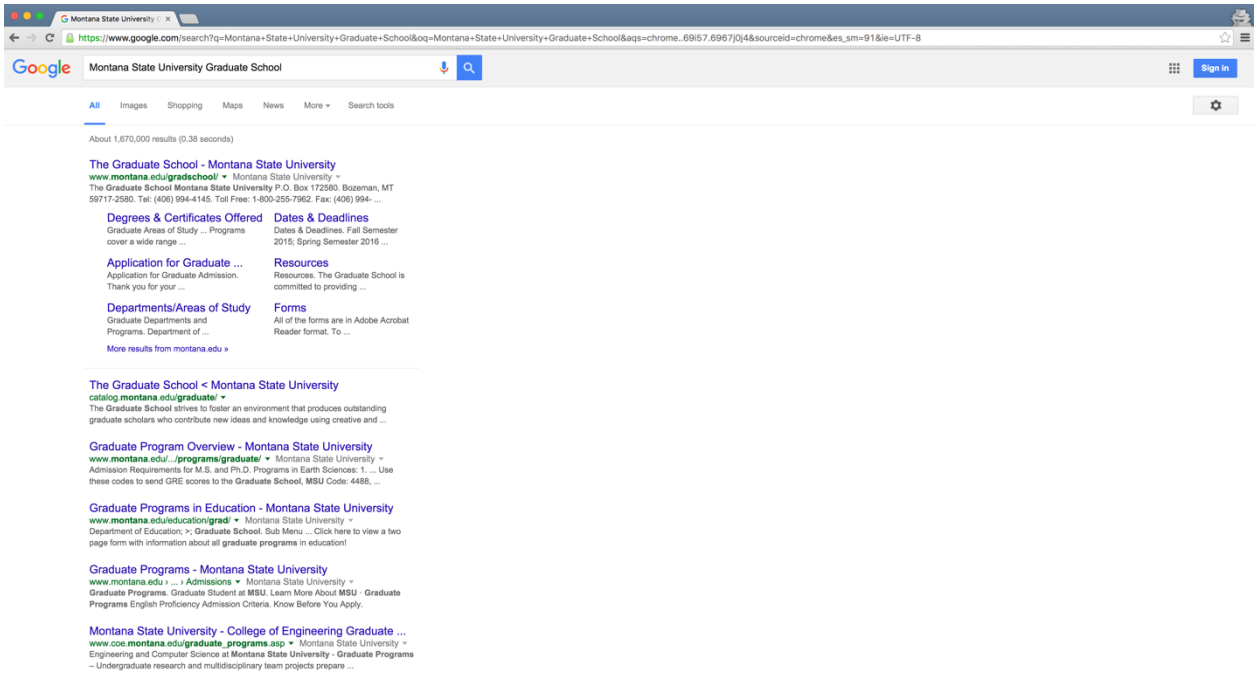

Figure 70: MSU Graduate School lacking a KC

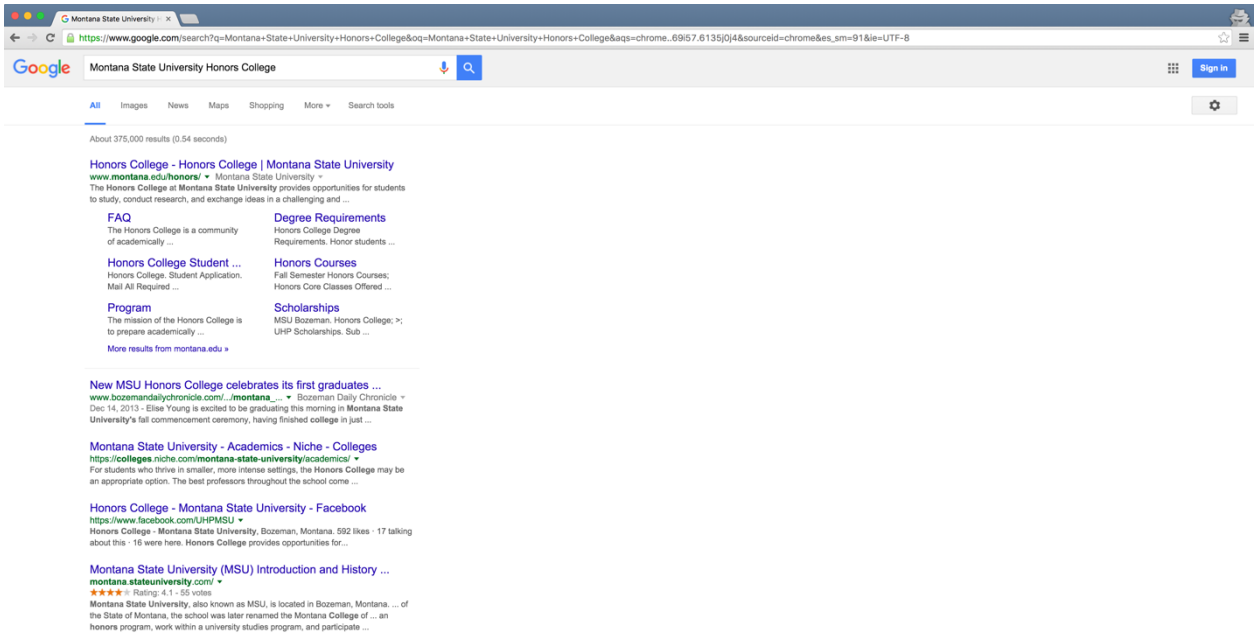

Figure 71: MSU Honors College lacking a KC prior to intervention by the MSU Library

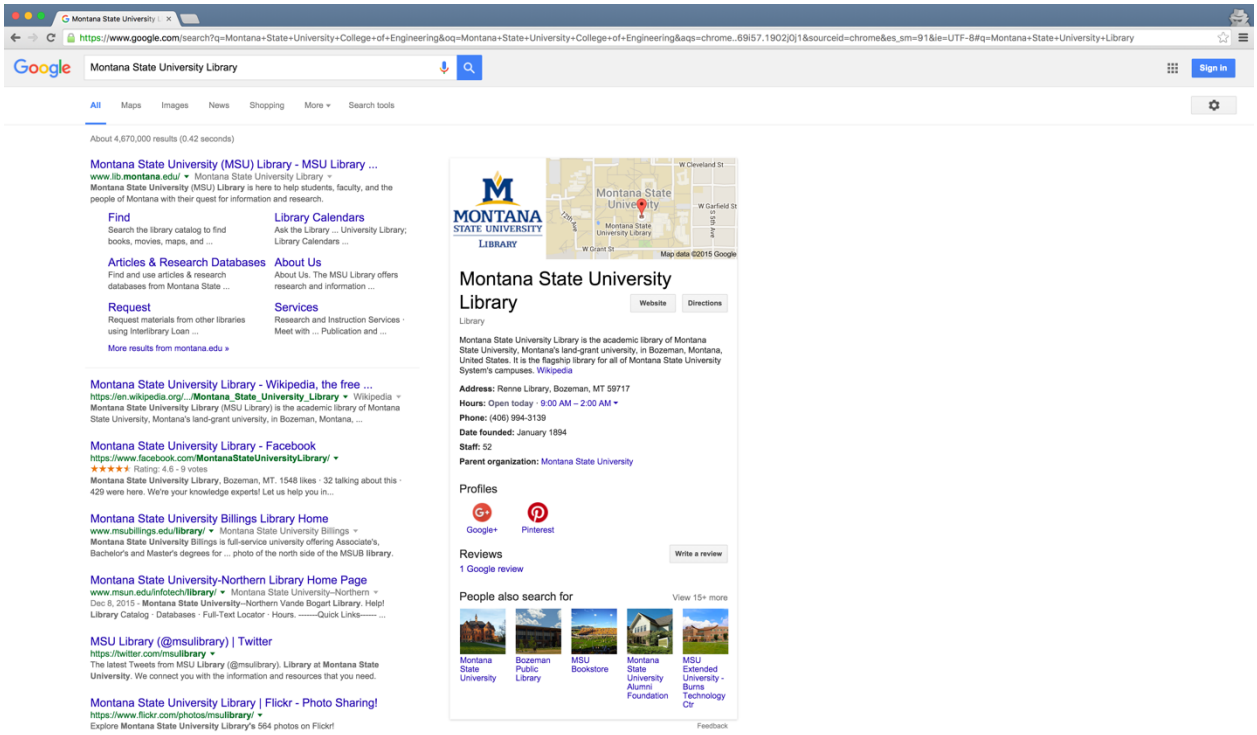

Figure 72: MSU Library, whose SWI had been established in 2014

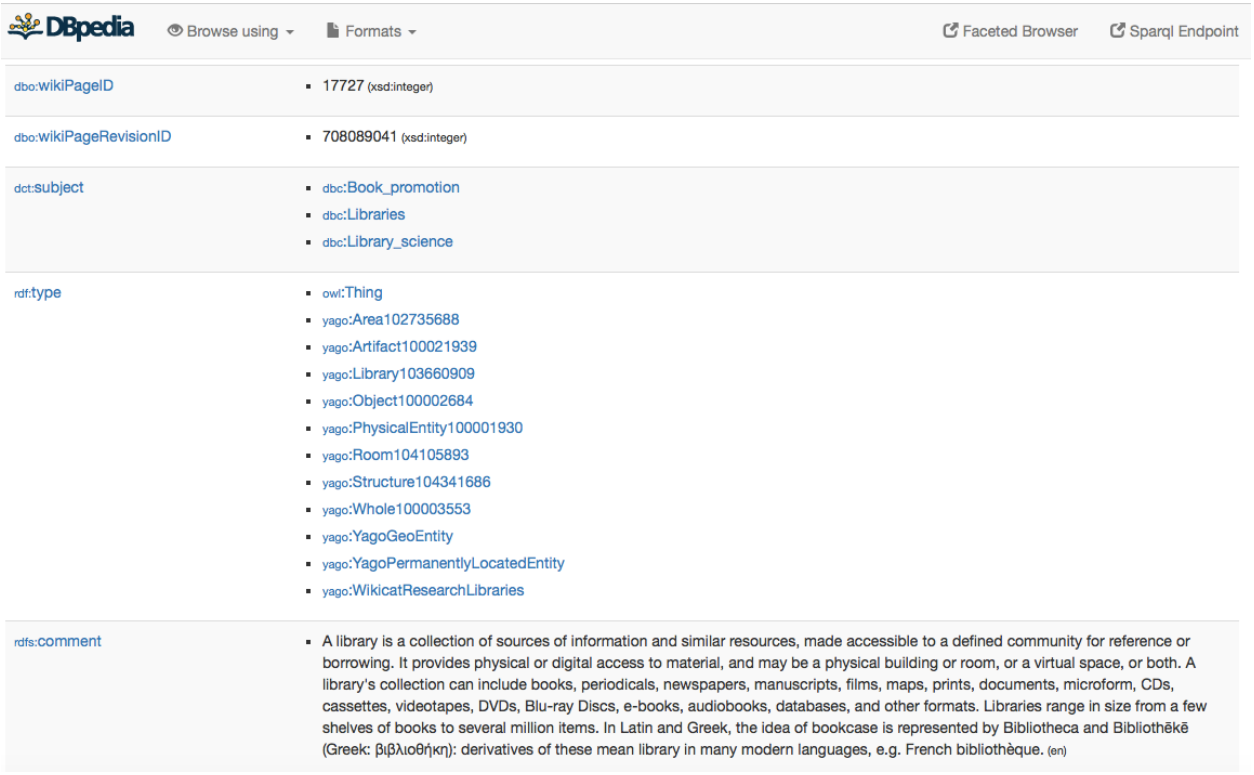

Figure 73: DBpedia record for "Library" from November 27, 2016

Appendix E: Data set readme file

```
This readme.txt file was generated on 2016-11-19 by Kenning Arlitsch
-------------------------------------------------
GENERAL INFORMATION
-------------------------------------------------
Title of Dataset: Data set supporting Ph.D. dissertation “Semantic Web
Identity in Academic Organizations: Search engine entity recognition and the
sources that influence Knowledge Graph Cards in search results”

Principal Investigator Contact Information
Name: Kenning Arlitsch
Institution: Montana State University
Address: P.O. Box 173320, MSU Library, Bozeman, MT 59717, USA
Email: kenning.arlitsch@montana.edu
Degree-granting institution: Institut für Bibliotheks- und
Informationswissenschaft (IBI)
Humboldt Universität zu Berlin
Address: Dorotheenstraße 26, Berlin, Germany

Date of data collection (single date, range, or approximate date): 2015-2016

Geographic location of data collection: Bozeman, MT 59717, USA

Date files were created: 2016

Are there multiple versions of the dataset? No

Information about funding sources that supported the collection of the data:
None



File Information:
Filename: Arlitsch-dissertation-dataset-metadata_2016-11-19.docx
Short description: Metadata required for submission of the dataset to Montana
State University ScholarWorks data repository
```

Filename: SWI-survey_2016-10-16.csv
Short description: Main spreadsheet containing recorded observations for 125 Association of Research Libraries (ARL) members. 125 primary names and 94 alternate names were searched for evidence of Knowledge Graph Cards (KC) in Google search results, and for evidence of records or articles in Google My Business, Google+, Wikipedia, DBpedia, and Wikidata.

Filename: SWI-survey-subset_2016-10-16.csv
Short description: This smaller spreadsheet was used to run statistical analysis in R for the parent institution of each of the 125 ARL member libraries, rather than the primary and alternate names of the libraries

Filename: SWI-analysis-final_2016-11-17.R
Short description: R source file with equations and commands used to analyze "SWI-survey spreadsheet" file.

Filename: SWI-subset-analysis-final_2016-11-17.R
Short description: R source file with equations and commands used to analyze "SWI-survey-subset" spreadsheet file.

Filename: SWI-DBpedia-screenshots.zip
Short description: Zipped archive containing 84 screen capture files in PNG format from DBpedia.

Filename: SWI-G+-screenshots.zip
Short description: Zipped archive containing 288 screen capture files in PNG format from Google+.

Filename: SWI-GMB-screenshots.zip
Short description: Zipped archive containing 179 screen capture files in PNG format from Google My Business.

Filename: SWI-Google-search-screenshots.zip
Short description: Zipped archive containing 245 screen capture files in PNG format from Google search results.

Filename: SWI-Wikidata-screenshots.zip
Short description: Zipped archive containing 230 screen capture files in PNG format from Wikidata.

Filename: SWI-Wikipedia-screenshots.zip
Short description: Zipped archive containing 223 screen capture files in PNG format from Wikipedia.

Filename: SWI-Wikidata-screenshots.zip
Short description: Zipped archive containing 230 screen capture files in PNG format from Wikidata.

Filename: SWI-MSU-Colleges-screenshots.zip
Short description: Zipped archive containing 72 screen capture files in PNG format from Google searches of eleven Montana State University colleges.

Filename: SWI-casestudy-CNI-screenshots.zip
Short description: Zipped archive containing 34 screen capture files in PNG format collected during case study development for the Coalition for Networked Information (CNI).

Filename: SWI-casestudy-McMaster-screenshots.zip
Short description: Zipped archive containing 21 screen capture files in PNG format and browser exports in PDF format, which were collected during case study development for McMaster University Libraries.

Filename: SWI-casestudy-MSU-library-screenshots.zip
Short description: Zipped archive containing 28 screen capture files in PNG format and browser exports in PDF format, which were collected during case study development for Montana State University Library.

If data set includes multiple files related to one another, include relationship here:
Screenshot files support the data recorded in the spreadsheet files. R source files contain statistical analysis commands and equations that were used to analyze the spreadsheet data.

METHODOLOGICAL INFORMATION

Description of methods used for collection/generation of data:
The Action Research methodology guided this research. Data collection methods included screen captures of search results conducted in Google, Google My Business, Google+, Wikipedia, DBpedia, and Wikidata. Results of searches were also recorded in two spreadsheets. The Chrome web browser was used in Icognito mode for most searches. The Safari web browser was used for Google+ searches.

Methods for processing the data: The R statistical software was used to analyze the data. Two R source files are included in this package.

Instrument-specific information needed to interpret the data: None

Standards and calibration information, if appropriate: None

Environmental/experimental conditions: None

Describe any quality-assurance procedures performed on the data:
Data integrity checks were conducted with R to find and correct spreadsheet errors. Errors were checked against screen capture files and spreadsheets notations were adjusted accordingly.

Codes or symbols used to note or characterize low quality/questionable outliers that people should be aware of:
Code/symbol: None
Definition: None

People involved with sample collection, processing, analysis and/or submission:
None

DATA-SPECIFIC INFORMATION

The following information applies to the two spreadsheet files included with this dataset.

Column headings for tabular data: PrimORAltKC
Full name: Primary or Alternate Knowledge Graph Card

Definition: Google Knowledge Graph Card appeared in search results for primary or alternate names of ARL libraries.
Units of measurement: Binary. Yes=1, No=0

Column headings for tabular data: ParentInstitution
Full name: Parent Institution
Definition: Name of the university or parent institution to which the ARL library belongs.

Column headings for tabular data: ARL Library Name
Full name: Association of Research Libraries Library Name
Definition: The primary and alternate name (where an alternate name exists) of the ARL member library. The primary name is derived from the ARL membership directory (http://www.arl.org/membership/list-of-arl-members) and is the official name submitted by the library organizations.

Column headings for tabular data: Primary
Full name: Primary
Definition: Column indicates in binary format (1,0) which of the names in the ARL Library Name column is defined as the primary (official) name of the library organization, as listed in the ARL membership directory (http://www.arl.org/membership/list-of-arl-members). A value of 1 indicates that the row contains the primary name; a value of 0 indicates the row does not contain the primary name.

Column headings for tabular data: KC
Full name: Knowledge Graph Card
Definition: Column indicates whether a Google Knowledge Graph card appeared in the search results for the name of the library being searched. 0 indicates no KC was found; 1 indicates a KC was found.

Column headings for tabular data: GMB
Full name: Google My Business
Definition: Knowledge base searched to determine whether a business had been claimed and verified for the primary or alternate name of the ARL library. 0 indicates no claimed and verified record could be found; 1 indicates a claimed and verified record was found.

Column headings for tabular data: Gplus

Full name: Google+ or Google Plus.
Definition: Name of the knowledge base that was searched to determine whether a verified or unverified profile existed for the primary or alternate name of the library organization. In this column, 0 indicates no profile was found, 1 indicates an unverified profile was found; 2 indicates a verified profile was found.

Column headings for tabular data: Wikipedia
Full name: Wikipedia — the Free Encyclopedia
Definition: Name of the knowledge base that was searched to determine whether an article had been published for the primary or alternate name of the library organization. 0 indicates no article was found; 1 indicates an article was found.

Column headings for tabular data: WikipediaInfobox
Full name: Wikipedia Infobox
Definition: This column recorded whether a Wikipedia article existed for the primary or alternate name of the library organization being searched, and whether the article (if found) included an infobox. 0 indicates no article was found; 1 indicates an article without infobox was found; 2 indicates an article with infobox was found.

Column headings for tabular data: DBpedia
Full name: DBpedia
Definition: Knowledge base that was searched to determine whether a structured data record had been generated from Wikipedia for the primary or alternate name of the library organization. This search was conducted on the dataset last made available by DBpedia in the spring of 2015. 0 indicates no record was found; 1 indicates a record was found.

Column headings for tabular data: Wikidata
Full name: Wikidata
Definition: Knowledge base that was searched to determine whether a structured data record existed for the primary or alternate name of the library organization. Records that contained fewer than two populated fields were not considered viable records. 0 indicates no record was found; 1 indicates a record was found.

Column headings for tabular data: AccurateKC

Full name: Accurate Knowledge Graph Card
Definition: This column indicates whether the KC that displayed for the primary or alternate name of the library was accurate for the library organization being searched. 0 indicates the KC was inaccurate; 1 indicates it was accurate.

Column headings for tabular data: AccurateKCInst
Full name: Accurate Knowledge Graph Card for the Institution
Definition: The Google Knowledge Graph Card that appeared in search results was accurate for the parent institution of the library organization being searched. 0 indicates the KC was inaccurate; 1 indicates it was accurate.

Column headings for tabular data: SameAs
Full name: Same As
Definition: When Google Knowledge Graph Cards appeared for both primary and alternate library names being searched, it was the same card that appeared, indicating that Google has a semantic understanding of the relationship of the two names to the same organization. 0 indicates that a different KC appeared for primary and alternate names; 1 indicates the same KC appeared whether the primary or alternate names were searched.

Column headings for tabular data: Logo
Full name: Logo or Map
Definition: This column captured whether a logo appeared in the KC as an information element. 0 indicates no logo appeared; 1 indicates a logo appeared.

Column headings for tabular data: Img
Full name: Image or Photograph
Definition: This column captured whether an image or photograph appeared in the KC as an information element. 0 indicates no image appeared; 1 indicates an image appeared.

Column headings for tabular data: Type
Full name: Type of organization
Definition: This column captured whether the type of organization was indicated in the KC as an information element. 0 no organization type was indicated; 1 an organization type was indicated.

Column headings for tabular data: Appearance
Full name: Appearance grouping
Definition: This column categorized the information elements Logo, Img, and Type as a single group. The value for each row in the Appearance column was calculated as a product of the three variables. If any of the variables had indicated a 0 then the entire Appearance group for that name was also recorded as a 0. This grouping was created because it was observed that these three variables almost always appeared together, i.e. if one appeared then it was rare for the other two to not appear.

Column headings for tabular data: Address
Full name: Physical address of the organization
Definition: This column captured whether an address for the library organization appeared in the KC. 0 indicates no address appeared; 1 indicates an address appeared.

Column headings for tabular data: Phone
Full name: Telephone number
Definition: This column captured whether telephone number for the library organization appeared in the KC. 0 indicates no phone number appeared; 1 indicates a phone number appeared.

Column headings for tabular data: Directions
Full name: Clickable button for directions to the physical address, provided by Google Maps.
Definition: This column captured whether a clickable button appeared in the KC that linked to directions to the library organization in Google Maps. 0 indicates no button appeared; 1 indicates a button appeared.

Column headings for tabular data: Website
Full name: Clickable button for the website
Definition: This column captured whether a clickable button appeared in the KC that linked to the library organization’s website. 0 indicates no button appeared; 1 indicates a button appeared.

Column headings for tabular data: Contact
Full name: Contact grouping
Definition: This column categorized the prior four information elements (Address, Phone, Directions, Website) as a single group. The value for each

row in the Contact column was calculated as a product of the four variables. If any of the variables had indicated a 0 then the entire Contact group for that name was also recorded as a 0. This grouping was created because it was observed that these three variables almost always appeared together, i.e. if one appeared then it was rare for the other two to not appear.

Column headings for tabular data: Hours
Full name: Operating hours of the library organization
Definition: This column captured whether the operating hours appeared in the KC for the library organization. 0 indicates no button appeared; 1 indicates a button appeared. While this information was collected, it was discarded from the statistical analysis because the appearance of hours on the KC was too variable and thus did not seem to fit with the Contact group.

Column headings for tabular data: Description
Full name: Textual description field on the KC
Definition: This column captured whether a brief textual description about the library organization appeared on the KC. 0 indicates no description appeared; 1 indicates a description appeared. This information element became a group of one as Google explicitly indicates its source as Wikipedia.

Column headings for tabular data: Comment
Full name: Comment
Definition: This column captured free text notes and observations made during data collection.

SHARING/ACCESS INFORMATION

Licenses/restrictions placed on the data: CC BY 4.0
https://creativecommons.org/licenses/by/4.0/

This data set is published from the United States.

CREDITS

Based on a template by University of Minnesota Libraries:
http://lib.umn.edu/datamanagement